Forecasting Failure
Assessing Risks to Quality Assurance in Higher Education Using Machine Learning

Griffiths, Alexander

Awarding institution:
King's College London

School of Management & Business

Forecasting Failure: Assessing Risks to Quality Assurance in Higher Education Using Machine Learning

Submitted by Alex Griffiths

A candidate for the Degree of Doctor of Philosophy 2016

Abstact

The landscape of UK higher education has changed significantly in the last five years. A tripling of tuition fees, the uncapping of student numbers, and an explosion in the number of 'alternative providers' typify a more marketised higher education sector (Brown and Carasso, 2013). With more providers than ever before competing for students, many with little experience and profit-driven motives, there is a clear danger that quality will suffer.

Faced with limited resource and an expanding, fiercely independent sector, the Government sought to protect quality by asking the Quality Assurance Agency for Higher Education (QAA) to adopt a risk-based approach. The 2011 White Paper *Student at the Heart of the System* directed QAA to prioritise their reviews based on "an objective assessment of a basket of data, monitored continually but at arm's length" (BIS, 2011, 3.19). There is, however, an evident dearth of empirical evidence to support such an approach . The aim of this thesis is to examine the extent to which available data can predict the outcome of quality assurance reviews, and hence prioritise them.

To fulfill this aim, the outcomes of all QAA reviews comparable with its current inspection methods were gathered along with all available data that could feasibly form part of a data-driven risk-based approach to quality assurance. Using machine learning, this study shows conclusively that a risk-based approach to quality assurance, as envisioned in the 2011 White Paper, cannot work. There is no connection between the available data and the subsequent outcome of QAA reviews.

The final part of this thesis therefore examines the reason *why* there is no connection between the available data and the outcome of QAA reivews. Three overarching and non-exclusive possibilities are identified. Concerns over the data, the review process, and the definition of 'quality' pose significant barriers to the operation of a successful data-driven, risk-based approach. An alternative approach to prioritising quality assurance in higher education is therefore required.

Ackowledgements

No thesis can be written without the significant help and support of countless actors. This thesis is no exception.

First and foremost, I am extremely grateful to Professor Alison Wolf CBE (Baroness Wolf of Dulwich) and Dr Henry Rothstein who have been my first and second supervisors respectively. Alison's drive and hard work resulted in the successful application for a collaborative Economic and Social Research Council (ESRC) studentship with the Quality Assurance Agency (QAA), for which I was fortunate enough to be selected. Her support, insight and ability to challenge accepted wisdom has proved vital in the development of this thesis, as has her championing of the findings greatly aided its impact. I am indebted to Henry for supervising the thesis despite his substantial workload over the three years and affording it such significant attention. His expertise in risk and regulation, and the many hours we spent musing the nature of 'quality', have proved vital in the development of the thesis.

Second, this study would not have been possible without the financial and, in the case of the latter, guidance and collaboration of the ESRC and QAA. Having worked in the public sector for many years I have seen a great deal of money spent on consultants with the aim of resolving regulatory problems quickly. More often that not these quick answers do little to resolve complex issues and merely serve as a metaphorical sticking plaster. It was refreshing therefore to see the QAA willing to engage in the necessarily slow process of collaborating with a PhD candidate in the hope of arriving at a meaningful solution, and also that QAA were willing to take the risk that, as has been the case, the thesis has not produced the result that they had hoped for. I am especially grateful to Dr. Liz Halford, the recently departed Director of Research at QAA who helped establish the study, and offered invaluable insight and support throughout the three years. My thanks also go to every other member of staff at QAA who has supported and engaged with me over the three years, the latter part of which was a challenging time for the organisation.

Third, I am extremely grateful for the support of Professor Harvey Goldstein, recipient of the Royal Statistical Society's Guy Medal and Fellow of the British Academy, whose statistical advice has been reassuring and invaluable.

Finally, I owe a great deal of gratitude to everyone who has tolerated my complete lack of availability or capacity to think too deeply about anything other than risk-based regulation and quality since 2013.

Contents

List of Acronyms

AAU	Academic Audit Unit
AoC	Association of Colleges
ARROW	Advanced Risk Responsive Operating Framework
AUC	Area Under the Curve
BBC	British Broadcasting Corporation
BERR	Department for Business, Enterprise and Regulatory Reform
BIS	Department for Business, Innovation & Skills
BRC	Better Regulation Commission
BRTF	Better Regulation Task Force
BTEC	British Technology Education Council
CDP	Committee of Directors of Polytechnics
CHES	Centre for Higher Education Studies
CNAA	Council for National Academic Awards
CQC	Care Quality Commission
CVCP	Committee of Vice Chancellors and Principals
DELNI	Department for Employment and Learning (Northern Ireland)
DES	Department for Education and Skills
DfE	Department for Education
DfEE	Department for Education and Employment
DLHE	Destinations of leavers from Higher Education
ECJ	European Court of Justice
ENQA	European Association for Quality Assurance in Higher Education
EPA	Environmental Protection Agency
EQAF	European Quality Assurance Forum
ESG	Standards and Guidelines for Quality Assurance in the European Higher Education Area
FEC	Further Education College
FPE	Full-Person Equivalent
FSA	Financial Services Authority
FSMG	Financial Sustainability, Management and Governance
FTE	Full-Time Equivalent
GOsC	General Osteopathic Council
HE in FE	Higher Education in Further Education
HEC	Higher Education Commission
HEFCE	Higher Education Funding Council for England
HEFCW	Higher Education Funding Council for Wales
HEI	Higher Education Institution
HEIDI	Higher Education Information Database for Institutions
HEIFES	Higher Education in Further Education Students
HEQC	Higher Education Quality Council
HER	Higher Education Review
HESA	Higher Education Statistics Agency
HESES	Higher Education Student Early Statistics
HMI	Her Majesty's Inspectorate
HNC	Higher National Certificate

HND	Higher National Diploma
HTS	Highly Trusted Status
ILR	Individual Learner's Record
IRS	Internal Revenue Service
IUSSC	Innovation, Universities, Science and Skills Committee
JPG	Joint Planning Group
JPIWG	Joint Performance Indicator Working Group
KFI	Key Financial Indicator
KNN	K-Nearest Neighbours
NAO	National Audit Office
NCIHE	National Committee of Inquiry into Higher Education
NSS	National Student Survey
OECD	Organisation for Economic Co-operation and Development
OFFA	Office for Fair Access
OfS	Office for Students
Ofsted	Office for Standards in Education, Children's Services and Skills
PAC	Public Accounts Committee
PCFC	Polytechnics and Colleges Funding Council
QAA	Quality Assurance Agency for Higher Education
QARSG	Quality Assessment Review Steering Group
RCGP	Royal College of General Practitioners
REO	Review of Educational Oversight
ROC	Receiver Operating Characteristic
RSCD	Review for Specific Course Designation
SCOP	Standing Conference of Principals
SFA	Skills Funding Agency
SLC	Student Loans Company
SRA	Solicitors Regulation Authority
SRHE	Society for Research into Higher Education
TEF	Teaching Excellence Framework
TEQSA	Tertiary Education Quality Standards Agency
THE	Times Higher Education
TQA	Teaching Quality Assessment
UCAS	University and College Admission Service
UFC	Universities Funding Council
UGC	University Grants Committee

List of Figures

List of Tables

1. Introduction

This introduction briefly explains how this thesis came into being, its aim, and the approach adopted to achieve this aim. It begins by detailing the developments in higher education policy and regulatory theory that led to the conception of the thesis, the author's background and interest in the topic, and the impact of the findings to date. This is followed by a chapter-by-chapter overview of the thesis.

1.1. Thesis Development and Aims

In 2012, the UK higher education landscape was undergoing significant change. New undergraduate students would be the first to face fees of up to £9,000 a year, students attending new, for-profit 'alternative providers' would be allowed access to Government-backed student loans, and the restrictions on the use of the protected term 'university' were being lessened. A sector which had been relatively stable, albeit expanding, for many years faced unprecendented change and 'marketisation' (Brown and Carasso, 2013). To keep pace with these developments and remain fit for purpose, the regulation of higher education also needed to change.

The Quality Assurance Agency for Higher Education (QAA) is "the independent body entrusted with monitoring, and advising on, standards and quality in UK higher education" (QAA, 2015a). Since 2002, QAA had been operating a largely-unchanged system of sexennial audits of providers' internal quality assurance processes (QAA, 2002). The 2011 White Paper *Students at the Heart of the System* called for the QAA to adopt a new:

> "…genuinely risk-based approach, focusing QAA effort where it will have most impact and giving students power to hold universities to account. All providers must continue to be part of a single assurance framework. But we would explore options in which the frequency – and perhaps need – for a full, scheduled institutional review will depend on an objective assessment of a basket of data, monitored continually but at arm's length."
>
> (BIS, 2011, 3.19)

'Risk-based' approaches to regulation became popular at the start of the 21st century and are defined by the explicit allocation of regulatory resource in proportion to the risks posed to the regulator's objectives (Black, 2005; Rothstein *et al.*, 2006b). The enthusiasm for risk-based approaches resulted in their use becoming a statutory requirement for all regulators in 2008 (BERR, 2007). Up until 2012, QAA had managed to avoid this statutory requirement by virtue of their unique status as a private company contracted by regulators – the national funding councils

- to deliver part of their regulatory function (Further and Higher Education Act, 1992; Brown, 2004). Despite the fact that regulators had been practising 'risk-based' approaches for up to ten years by this point, there was an evident dearth of empirical evidence to inform such a policy (Raban, 2011).

Conversations between Professor Alison Wolf, Sir Roy Griffiths Chair of Public Sector Management at King's College London, and Anthony McClaren, then Chief Executive of QAA, concerning this lack of evidence resulted in this collaborative ESRC and QAA-funded thesis; the idea behind which is to provide that missing evidence. More specifically, the aim of this thesis is to examine the extent to which available data could predict the outcome of QAA reviews, and hence prioritise them, as part of a risk-based approach to quality assurance.

Since the publication of *Students at the Heart of the System*, and the subsequent conception of this thesis, there has been significant change in higher education oversight. HEFCE have redesigned the quality assurance system, divided the work into six separate contracts, and awarded four of those contracts to QAA following a competitive tender process (HEFCE, 2016c). The Conservative Party won a majority in the 2015 General Election, published green and white higher education papers, and are now taking the *Higher Education and Research Bill* through parliament (BIS, 2015b, 2016; Department for Education, 2016). The proposed legislative changes, which may lead to further changes to quality assurance and the introduction of a Teaching Excellence Framework (TEF), prompted a Business, Innovation and Skills (BIS) select committee hearing into 'Assessing Quality in Higher Education' (BIS Select Committee, 2015, 2016a). BIS has since been replaced by the Department for Business, Energy & Industrial Strategy and responibility for higher education has moved to the Department for Education (DfE) following the post EU-referendum reconfiguration of Whitehall (Leach, 2016).

This thesis, however, has a specific and stable line of enquiry. Whilst I will allude to the aforementioned events and developments at various points during the course of my research, I have neither set out to evaluate current proposals for reform, nor track the changing politics of higher education regulation. This thesis is about the extent to which available data could predict the outcome of QAA reviews, and hence prioritise them. It is not about today's, or yesterday's, particular item of debate. It's about whether QAA judgements can be predicted using a basket of available data, and what one can conclude about risk-based approaches in higher education as a result of the answer to this question.

At the time I applied for the studentship in July 2013, I had been working for four-and-a-half years in the 'Intelligence' Directorate of the Care Quality Commission (CQC), England's health and social

care regulator. There, my role was to lead a team of analysts in the development and automation of 'Quality and Risk Profiles' (QRPs). These automated risk reports could contain in excess of 1,000 quantitative and qualitative metrics and required a vast amount of resource to maintain. Despite this, they appeared to be of little use in prioritsing inspections (see for example Pollard, 2011; Francis, 2013). Rather than continue to implement what I saw as a flawed approach, I was keen to apply my skills to the challenge of identifying an effective, evidence-based approach in another sector.

This thesis is the result. Somewhat dishearteningly for both QAA and for myself, having spent my career to date working with regulatory performance data, this thesis conclusively demonstrates a data-driven, risk-based approach to quality assurance in higher education cannot work. There is simply no robust relationship between the vast array of UK higher education data and the subsequent outcome of QAA reviews. These quantitative findings resulted in the final part of the thesis that explores *why* the available data is not able to predict the outcome of QAA reviews.

The quantitative results have contributed to a change in quality assurance policy. The findings were widely publicised in the quality assurance world in late 2015 following a specially convened seminar of the Policy Institute at King's, the author's oral and written evidence to the BIS Select Committee's 'Assessing Quality in Higher Education' Inquiry, presentation at the 2015 Annual Conferences' of the European Quality Assurance Forum (EQAF) and the Society for Research into Higher Education (SRHE), and coverage in the *Times Higher Education* (Griffiths, 2015; EQAF, 2015; SRHE, 2015; Havergal, 2015). The use of expert interpretation of metrics is now being advocated in the unevidenced hope that it will allow for the successful prioritisation of QAA reviews (Kimber, 2015; HEFCE, 2016c).

This thesis therefore came about as the result of the 2011 White Paper which called for the QAA to adopt a data-led, risk-based approach to scheduling their reviews. The aim of this thesis is to examine the extent to which available data could predict the outcome of quality assurance reviews, and hence prioritise them. The approach taken to meet this aim is detailed below.

1.2. Thesis Outline

This thesis is presented in a sequential fashion. The first three chapters introduce the thesis and establish the research problem via an examination of the development of quality assurance in higher education and risk-based regulation. Chapters four to seven detail the data and methods

used for the quantitative analysis, and the subsequent results for each of the three provider types in higher education. Finally, chapters eight and nine explore the possible reasons for the results of the quantitative analysis, and discuss the implications of the overall findings. The specific chapter structure is as follows:

- Chapter 1: Introduction
- Chapter 2: The Evolution of Quality Assurance in UK Higher Education.

  This chapter details the development of quality assurance in UK higher education. It covers the introduction of formal quality assurance and assessment, the creation of the QAA to oversee a single quality system, external factors shaping the QAA's approach, and the introduction of a risk-based approach to quality assurance.

- Chapter 3: Risk-Based Approaches

  This chapter explores 'risk-based' approaches to regulation. It defines 'risk-based regulation' and examines its evolution, merits, and limitations. This is followed by an examination of the different ways in which a risk-based approach can be implemented and a review of the literature concerning risk-based approaches to quality assurance in higher education.

- Chapter 4: Data and Methods

  This chapter defines the research question and details the data available to answer it, how that data was prepared, and the modelling and evaluation approaches chosen.

- Chapter 5: Predicting the Outcome of Higher Education Insititution (HEI) Reviews

  This chapter determines which metrics, if any, could have predicted the outcome of past QAA HEI reviews as part of a robust model and how accurately they could have done so. It details the specific challenges of the HEI subsector before peforming four analyses using naturally-complete, imputed, in-year standardised, and benchmarked data.

- Chapter 6: Predicting the Outcome of Further Education College (FEC) Reviews

  Similar to chapter five, this chapter determines which metrics, if any, could have predicted the outcome of past QAA FEC reviews as part of a robust model and how accurately they could have done so. It details the specific challenges of the FEC subsector before performing two analyses using naturally-complete and in-year standardised data.

- Chapter 7: Predicting the Outcome of Alternative Provider Reviews

Again, similar to chapters five and six, this chapter determines which metrics, if any, could have predicted the outcome of past QAA reviews of alternative providers, and how accurately they could have done so. The specific challenges of the alternative provider subsector are explored before four analyses are conducted. The greater number and diversity of review judgements available for alternative providers allow for question-level analyses to be performed in addition to the overall review-level analysis explored in the previous chapters.

- Chapter 8: Why Can't the Available Data Predict the Outcome of QAA Reviews?

  Following on from the quantitative chapters that concluded that a data-driven, risk-based approach to quality assurance in higher education cannot work, this chapter aims to identify the reasons *why* the available data cannot predict the outcome of QAA reviews. The chapter identifies three overarching reasons: issues with the available data, issues with QAA reviews, and the subjective, contested nature of quality.

- Chapter 9. Discussion

  The final chapter discusses the findings of this thesis. It addresses the conclusions that can be drawn from the thesis and their potential limitations; the next steps for the quality assurance of higher education; and the meaning of the findings in the wider context of data-driven, risk-based regulation outside of higher education.

The thesis now continues with a review of the development of quality assurance in UK higher education.

2. The Evolution of Quality Assurance in UK Higher Education

"The Government, on behalf of tax payers, has a legitimate interest in knowing that the public funding which goes into higher education is well spent in supporting the UK economy and society."

(QAA, 2010, p.2)

The purpose of this chapter is to detail the development of quality assurance in UK higher education. It will cover the introduction of formal quality assurance and assessment, the creation of the QAA to oversee a single quality system, external factors shaping the QAA's approach, and the introduction of a risk-based approach to quality assurance. A Timeline of key events is available in Appendix A at the end of the chapter.

2.1. The Introduction of Sector-Wide External Quality Assurance

With the goal of improved economic performance, each UK Government since the publication of the Robbins report in 1963 has successfully sought the expansion of the higher education sector. From 113,000 students in the system in 1961/2, numbers have grown to 909,300 in 1985/6, 1,408,800 in 1992/93 and 2,496,645 in 2011/12: a twenty-two fold increase in just over fifty years (HESA, 2013). As the number of students has grown however, so too has the amount of public resource required to fund them and the need to ensure that such funds are used effectively. Following a period of economic decline, the 1980/81 – 1983/84 spending round saw a 15% cut in the funding for higher education coupled with a continuing expansion of student numbers. Under such financial pressure it was clear that quality could suffer. In 1983 the Society for Research into Higher Education reported that higher education institutions:

> "may in the future find themselves under pressure to compromise academic quality in attempting to maintain student numbers or earn income from other sources. Some coordination of arrangements is desirable to try and ensure, for example, that academic standards for similar activities do not diverge too widely between institutions"

(SRHE, 1983, p.14-15)

The concern was shared by the Secretary of State, Sir Keith Joseph, who in the same year wrote to the University Grants Committee (UGC) asking them to assess how standards were currently being maintained and enhanced and, in the context of a more efficient use of resources in universities, to explore the possibilities for maintaining and improving academic quality in the

future. This led to the establishment of the 'Reynolds Group' by The Committee of Vice Chancellors and Principals (CVCP, now Universities UK) to "study and report on institutions' methods for maintaining and monitoring academic quality and standards" (CVCP, 1986, p.10). The group's report contained three formal codes of practice in addition to 'points of reference' for institutions undertaking self-comparisons concerning the maintenance and monitoring of standards. Following further encouragement from a new Secretary of State - Kenneth Baker - this in turn led to the CVCP's somewhat reluctant establishment of the self-regulatory Academic Audit Unit (AAU), designed to provide a more permanent oversight of universities' quality control systems, which commenced operations in 1990. Notwithstanding institutionally-appointed external examiners and certain professional areas, the AAU provided the first 'external' oversight of UK universities in their history (Brown and Carasso, 2013).

This contrasted strongly with the more rigorous quality measures for polytechnics and colleges (the 'public sector') where degrees and diplomas were awarded not by the institutions, but by the Council for National Academic Awards (CNAA) and the British Technology Education Council (BTEC) respectively. CNAA was established in 1964 to assure the quality of degrees in the public sector which originally could only be awarded following close scrutiny of individual courses. Scrutiny requirements were relaxed over time, however, with institutions validating their own courses once they had earned accreditation from CNAA (Harris, 1990). This oversight was coupled with formal inspection at both subject and institution-level by Her Majesty's Inspectorate (HMI) not dissimilar to its inspections of schools (Brown, 2004).

The 1988 White Paper established the Polytechnics and Colleges Funding Council (PCFC) and the Universities Funding Council (UFC), which would no longer fund institutions through grants but rather would operate a system of contracting. Ostensibly formed to drive the sustained expansion of higher education and the continued 'efficiency gains' necessary to fund it, the PCFC 'nationalised' the polytechnics, removing ownership from local authorities, as the universities were similarly being removed from the protected autonomy of the UGC (Jenkins, 1995). The Government's stated intention was to:

- encourage institutions to be more enterprising in attracting contracts from other sources, particularly in the private sector, and thereby to lessen their present degree of funding on public funding;
- sharpen accountability for the use of the public funds which will continue to be required;

- strengthen the commitment of institutions to the delivery of the educational services which it is agreed with the new planning and funding body they should provide.

(DES, 1987, paragraph 4.17)

Such enterprise was encouraged in 1990 by a cost neutral increase in tuition fees and decrease in block grant funding, both ultimately paid for by Government for UK nationals. It was now more lucrative to attract additional EU students as well as non-EU students who had been full fee paying since 1979. The unfettered expansion of higher education was slowed three years later however when limits were imposed on the number of funded places available at each institution.

The 1991 White Paper *Higher Education: A New Framework* (DES, 1991) detailed the Government's plans for a still larger and more cost-effective higher education sector, responsive to the needs of industry, to continue to support the drive for economic growth. The White Paper declared that UK higher education was "more efficient and more effective" than ever before and that recent growth, seeing one in five of all 18-19 year olds now entering higher education compared to one in seven in 1987, had occurred without a decline in quality evidenced by a steady increase in the proportion of first and second class degrees awarded (DES, 1991, p.3). The claimed maintenance of quality was allegedly possible due to the utilisation of capacity 'at the margin', mostly in the 'public sector' institutions which, at the cost of a drastically reduced unit of funding (Watson and Bowden, 1997), now educated more students than the university sector (DES, 1991).

Others, however, would strongly disagree with that analysis. Stevens (2005) claims there was a gradual decline in standards since 1970 that was not slowed until the Blair reforms of 2001. Only 28% of Vice-chancellors surveyed in 1993 felt that degree standards were being maintained (THE, 1993a). The reliance on the number of first and second class degrees awarded as a measure of quality ignored the possibility of grade inflation incentivised by increased competition, a continued pattern which drew the ire of the House of Commons Innovation, Universities, Science and Skills Committee 18 years later (IUSSC, 2009b).

Regardless of the contemporary state of higher education, if the expansion was to continue without quality suffering, further efficiency gains would be needed. Increasing competition and the marketisation of the sector was seen as the best way to deliver this. In keeping with the wider Government policy at the time, the White Paper stated that:

> "The Government believes that the real key to achieving cost effective expansion lies in greater competition for funds and students. That can be best achieved by breaking down

> the increasingly artificial and unhelpful barriers between the universities, and the polytechnics and colleges."
>
> (DES, 1991, p.12)

There was a two-fold strategy to achieve this end. *First*, 'public sector' institutions were now, under certain criteria, able to adopt the title of 'university' and award their own degrees in place of the CNAA, which was to be disbanded. *Second*, national funding councils responsible for funding all higher education teaching in universities, polytechnics and colleges were to replace the PCFC and the UFC (DES, 1991).

Whilst the goal of an expanded, more cost-effective higher education sector had clear benefits, there were also clear risks to quality from increased student numbers, enhanced competition, a reduced 'unit of resource' cost, and structural changes. Moreover, as the sums of public money invested in the sector grew, so too did the demand for accountability. The White Paper continued:

> "The prime responsibility for maintaining and enhancing the quality of teaching and learning rests with each individual institution. At the same time, there is a need for proper accountability for the substantial public funds invested in higher education. As part of this, students and employers need improved information about quality if the full benefit of increased competition is to be obtained.
>
> As demand for higher education expands further, and as competition among institutions increases, as a result of the changes outlined in the preceding chapters, the Government considers that new arrangements for quality assurance in higher education will be required."
>
> (DES, 1991, p.24)

George Walden, a former Higher Education Minister, put the matter more candidly:

> "The danger of overstretching the capability of institutions to absorb extra students was not unforeseen when expansion was discussed in the mid-eighties. It would be a simple-minded Government who thought you could cram in more students indefinitely without risk to quality … Only market romantics seriously believed that, by a combination of belt tightening, better management, better teaching with modern methods, and attracting private money and foreign students, the value of a University degree in a hugely expanded system could continue unimpaired"
>
> (Walden, 1996, p.122)

The Government therefore had to determine to what extent the additional assurances provided in the 'public sector' would be amalgamated with the more laissez-faire approach adopted by universities. Both quality *audit*, defined as "external scrutiny aimed at providing guarantees that institutions have suitable quality control mechanisms in place", and quality *assessment*, defined as "external review of, and judgements about, the quality of teaching and learning in institutions", were deemed necessary to provide stakeholders, including the funding councils, with sufficient information concerning quality to ensure the full benefit of enhanced competition was realised (DES, 1991, p.24).

Recognising the importance higher education institutions attached to their academic freedom - in no small measure a result of the higher education sector's lobbying powers - the Government proposed the quality *audit* role should be carried out by a single quality assurance body owned by the institutions. However, the Government reserved the powers to establish a body of their own should institutions fail to do so satisfactorily. The Government believed any doubts over the effectiveness of the self-regulatory body would be offset by the institutions' self-interest in demonstrating the rigour of their internal controls; failure to do so would result in a loss of reputation in a competitive national and international higher education market and would lead to the statutory regulation of academic standards (DES, 1991).

More salient would be the work of the 'quality *assessment* units' within each funding council whose assessments of what is actually provided would result in the award of an institutional quality ranking of 'excellent', 'satisfactory' or 'unsatisfactory' and inform the Council's funding decisions. Quality was to be assessed via 'subject review' (what was to become known as the Teaching Quality Assessment (TQA)) in two ways. First, quantitative performance indicators and value-added calculations would provide an overview of institutions and their courses. Second, direct observation of teaching and learning, management and organisation, accommodation, and equipment by professional staff, predominantly recruited from HMI, would provide a more comprehensive view of quality (Brown, 2004).

Overall reactions to the White Paper from representative bodies were largely positive. In their formal response to the White Paper the CVCP also supported the abolition of the distinction between universities and polytechnics and colleges and welcomed the Government's invitation to discuss how a new quality assurance regime might best operate. The CVCP, however, was less enthusiastic about the quality *assessment* regime stating that "the prime responsibility for the maintenance and enhancement of quality must rest with the institutions, reinforced by the constructive criticism of the quality audit unit" (CVCP, 1991, p.2). The CVCP proposed alternative

was for assessment units composed in part of academics seconded for "up-to-date knowledge of their subject" (CVCP, 1991, p.5). Despite suggestions by a minority of members of the Committee of Directors of Polytechnics (CDP) that polytechnics should embrace the new quality regime as a means of demonstrating their superiority in teaching over the existing universities, largely unsuccessful attempts were made by the CDP to water down the quality assessment proposals (Brown, 2004).

Later in 1991, a working group of institutional heads from the representative bodies (CVCP, CDP and Standing Conference of Principals (SCOP)) was formed to determine proposals for the quality regime. The results, sent to the Department for Education and Skills (DES) in October, acknowledged the institutions' collective responsibility for adequate systems of quality assurance across the sector and the need for a strong quality assurance body to "provide both necessary public assurance of institutional quality and act as an additional input to a pluralistic assessment of quality in higher education" (CVCP *et al.*, 1991, p.2). Concerns were expressed, however, over the duplication between the proposed quality *audits* and *assessments* to be conducted by the UK-wide quality assurance body and national funding councils respectively. The working group felt strongly that, whilst there was a clear need for funding councils to have access to quality information, this should remain independent from the funding operation. Peter Knight, the vice-chancellor of the University of Central England, was more forthright in his criticism declaring the quality system as "the most fundamental threat to academic freedom that higher education has ever effected … The one place judgements about quality should never be is in the hands of the funders. It gives them control of body and soul" (THE, 1993b).

The representative bodies recommended a 'steering committee' populated in the majority by representatives from institutions to advise on and approve the audit unit's annual work programme. As discussions continued the *Further and Higher Education Act* (1992) received a relatively smooth parliamentary approval. In keeping with the White Paper, the assessment councils within each Funding Council were established by section 70 which stated that each Funding Council shall:

(a) Secure that provision is made for assessing the quality of education provided in institutions for whose activities they provide, or are considering providing, financial support under this part of this Act, and
(b) Establish a committee, to be known as the "Quality Assessment Committee", with the function of giving them advice on the discharge of their duty under paragraph (a) above and such other functions as may be conferred on the committee by the council

The need for quality *audit* was somewhat more opaque, stated in section 82:

> (2) Any two or more councils shall, if directed to do so by the Secretary of State, jointly make provision for the assessment by a person appointed by them of matters relating to the arrangements made by each institution in Great Britain which is in the higher education sector for maintaining academic standards in the institution.
>
> (Further and Higher Education Act, 1992)

Only minor objections were raised concerning quality arrangements, most notably Baroness Blackstone, at the time Master of Birkbeck College and later the Minister for Education, who felt that housing the quality assessment body in the Funding Councils was "deeply flawed" and proposed a single, more rigorous, independent, UK-wide body "with duties to undertake [both quality assessment and audit] work and report on it, rather than tagging on a responsibility for the funding councils to set up committees" (HL Deb, 1991). This was rejected at the time, although adopted five years later, on the grounds that the funding councils were best placed to make decisions on the allocation of resources if they had first-hand experience of the quality of the institutions they were awarding; moreover, the audit was a vital part of academic freedom and the HEIs should retain their power to assure themselves of their quality (HL Deb, 1991).

Following discussions between the representative higher education bodies and DES, the Higher Education Quality Council (HEQC) was incorporated in May 1992 as a limited company, owned by the representative bodies, as the quality assurance body for higher education in the United Kingdom. The Board of Directors constituted of an equal number of university and polytechnic heads, two college representatives and two, later increased to four, independent members. Membership of the body was a condition of receiving public funds from the Funding Councils and subscriptions were to be deducted up front from institution's allotted funds from their Funding Council (Brown, 2004).

In summary, a somewhat confusing, and to many unnecessary, dual quality assurance regime was developed comprising both quality *audit* and quality *assessment*. Quality *audit* was to involve peer review on institutions' quality control mechanisms and be run by the UK-wide HEQC. Quality *assessment* was to involve the direct observation of teaching by professional inspectors employed by the four national funding councils. For established universities, this would be their first significant external assessment. For former 'public sector' institutions used to the scrutiny of CNAA and HMI, the dual system was much less of a change.

2.2. Dissatisfaction with HEQC, Quality Assessment Councils, and the establishment of the Quality Assurance Agency (QAA)

The HEQC's initial approach to quality *audit* was a continuation of the Academic Audit Unit's approach; three unaccompanied peer-auditors would visit an institution for three days and use 'primary documentation' to focus on four main areas:

- the provision and design of course and degree programmes
- teaching and communication methods
- academic staff
- the means of taking account of external examiners' reports and the views of students' and external bodies'.

The resulting report, which reviewed quality arrangements not against a national standard but rather the institution's own goals, would remain the property of each university (CVCP, 1992). Subsequent incremental changes, including the extension of audits to include universities' collaborative and overseas provision, took place alongside the development of the 'Graduate Standards Programme' instigated by the Secretary of State's concerns over academic standards, in part raised by visits to Singapore and Malaysia in 1994 where he received a number of complaints over the practices of British institutions attempting to recruit international students. In conjunction with the CVCP, HEQC would establish a set of threshold, or minimum acceptable, standards to ensure 'broad comparability' within a diverse sector. The full Graduate Standards Programme would never be enacted however following QAA's succession of HEQC shortly after the consultation was completed.

From 1992-95 the Funding Councils' approach to *assessment* saw subject areas within universities declare whether their performance against their own aims and objectives (and not a universal standard or threshold) was 'excellent', 'satisfactory' or 'unsatisfactory'. Of the 972 self-assessments, 553 were followed-up by peers who found 251 instances of 'excellent' provision, 290 instances of 'satisfactory' provision, and just 12 instances of 'unsatisfactory' provision. The remaining 419 cases had their declaration of 'satisfactory' accepted without review (HEFCE, 1995). Overall, the assessment regime did little to differentiate the quality of provision: 251 subjects within providers were deemed 'excellent', 709 'satisfactory', and just 12 'unsatisfactory'.

At this stage it was noted that "there is a danger that the costs of the whole exercise to the system, both to the Funding Councils and to the universities, will exceed the funds affected by the outcome" (Wagner, 1993, 281). However, following the Secretary of State's concern over

standards in 1994, the process was reviewed and, in addition to value for money and quality improvement, a third purpose of providing accessible information on the quality of education was added. Following an evaluation of the *assessment* process by the Centre for Higher Education Studies (CHES, 1994), visits were extended to cover all institutions and departments and judgements of 'excellent', 'satisfactory' or 'unsatisfactory' were replaced by a score of 1 (the lowest) to 4 (the highest) for each of six core areas leading to a published overall profile of judgements containing a score out of 24.

Almost as soon as HEQC and the Funding Councils began their work however, concerns were being raised over the efficacy of, and burden created by, the dual quality *audit* and *assessment* regime. Later in 1992 Dr Malcom Frazer, Chief Executive of HEQC, stated it would be better "if quality was not in the hands of the funding council or the HEQC" and that, when the binary divide between universities and polytechnics and colleges ceased, he had hoped "we would have an independent accrediting body which would accredit at both institutional level and course level" (Brookman, 1992). In December the *Times Higher Education Supplement* started its 'Quality Debate' series with the article 'Quality Assurance Arrangements are Going Wrong' (THE, 1992). Articles such as 'VCs Slam Red Tape' reporting "a rising tide of protests by institutions who say they will be bogged down by bureaucracy" (THE, 1993c, p.3) represented the majority view captured in a phone survey of Vice-chancellors published later that month, which reported that 82% condemned the existing quality arrangements and 71% said there should be a single quality body (THE, 1993a). Assiduous criticism was focused at the potential burden resulting from the new dual system, especially from the quality *assessment* work undertaken by the funding councils and the perceived deluge of repetitious and overlapping paperwork. The lack of confidence in the system was perhaps best illustrated by the admission of the Chief Executive of the Higher Education Funding Council for England (HEFCE) that, once they published their first critical quality assessment, he expected the council to be taken to judicial review. Condemnation of the system was not universal however: 58% of Vice-chancellors agreed that quality *audit* and *assessment* were useful guides to progress whilst some argued that an effective quality assurance regime would promote good practice and reassure students and the public whilst imposing a minimal burden (THE, 1993a; Foster, 1993).

The specific objections to the 'subject review' *assessment* process undertaken by the national funding councils were well documented and numerous. It was alleged that the system imposed excessive demands on institutions with a substantial proportion of the costs of both audit and assessment being transferred from the Government on to providers (Watson, 1995). The burden of these exercises was far greater than universities had expected or experienced previously. Moreover, it was argued that academic freedom and autonomy were being violated; that a culture

of compliance and 'hard-managerialism' was developing leading to managerial intrusion into academic matters; emphasising presentational and procedural matters at the potential expense of intellectual substance*;* failure to differentiate enough between outcomes to justify the use of scarce resources; and harming Britain's well-earned reputation for quality (Wagner, 1993; Alderman, 1996; NCIHE, 1997; Brennan, 1997).

The work of the HEQC was seen to exemplify Power's 'Audit Society' (1997) in which quality *audits* were described as 'rituals of verification' decoupled from the day-to-day work of institutions. In higher education, as in many other sectors, 'Quality' is difficult to measure, yet in an era of rising accountability, it had to be measured (Pollitt, 1993, 1995). Hence, the object of audit in HEIs was not the difficult to assess teaching and learning activities, but the more verifiable systems and processes which supposedly control those activities (Watson, 1995; Brown, 2004). The perceived impact of performing poorly in an HEQC audit led institutions to establish units responsible for ensuring success (Power, 1997, p101-2). Whilst often successful at ensuring a positive audit with reams of evidence submitted, checklists completed and performance measured, such centralised units within HEIs could further the measurement of what was auditable rather than what was originally intended (Power, 1997; Shore and Wright, 1999).

The overarching criticism of the dual system of quality assurance was that it was not clear *who* the process was assuring about *what*. Brown (1997) contends that a*udit* in higher education is intended to assure institutions that their own procedures are working as intended. *Audit* assumes "quality improvement is most likely to be achieved through the 'internal' professional motivation to do better rather than 'extrinsic' motivators such as money or prestige" (Brown, 1997, p.5). Conversely, *assessment* is intended to reassure external stakeholders over value for money and is based on the fundamental assumption that "quality improvement is most likely to be achieved through the motivation to compete to win additional students and resources" (Brown, 1997, p.6). The parallel *audit* and *assessment* processes therefore had contrasting operating assumptions and motivations.

The impact of the quality assurance regime was not perceived entirely negatively. Clark (1994) noted that many departments were introducing quality assurance procedures with external assessors for the first time. Moreover, a substantial number of academics were exploring other departments and experiencing new methods of teaching. Watson (1995) also noted greater attention being paid to teaching and learning performance, professional development, and how the infrastructure of institutions could best meet the needs of their students. Finally, the quality

assurance regime provided a source of 'reliable' and independent information for staff and students (Gordon, 2002).

In spite of the quality assurance system's stated benefits, pressure for change built and in 1994 the new Secretary of State, Mrs Gillian Shephard, asked the HEFCE Chief Executive, Professor (now Sir) Graeme Davies, to work with the representative bodies to determine how audit and assessment could be combined to form a single quality system.

Shortly afterwards, in a speech to the CVCP Committee, Mrs Shephard detailed her requirements for the unified system. It must:

- Provide assurance that standards of degrees are maintained and are broadly comparable – which does not mean identical – and that the quality of teaching and learning is such that students have the best opportunities of reaching those standards.
- Be transparent so that it can assist in enabling choices to be made by: universities and colleges themselves in deploying their resources in full knowledge of their strengths and weaknesses; potential students about university and course; employers in recruiting graduates; and the Government and the Funding Council in deploying public funds.
- Respect academic autonomy whilst having an external element.
- Respect academic diversity and freedom while at the same time addressing value for money and public accountability.
- Encourage the enhancement of quality and dissemination of good practice.
- Be cost-effective and avoid unreasonable burdens on institutions.

(Shephard, 1995) reported in (Brown, 2004).

After separate proposals by HEFCE, CVCP and HEQC for a unified system were unsuccessful, the CVCP proposed the establishment of a Joint Planning Group (JPG) with HEFCE to further develop their own ideas. Gillian Shephard welcomed the proposals but struck a note of caution: she detailed that costs and benefits should be established and compared to the existing regime; the timeframe for subject reviews should not exceed two years to allow the Funding Council, potential students and employers to make comparisons; and "In respect of assessment at least, I could not contemplate a solution which relied mainly on self-regulation" (Shephard, 1995).

The JPG was convened in September 1995 and, just five months later, the National Committee of Inquiry into Higher Education (NCIHE) or the "Dearing Committee" was established by the Secretary of State to "make recommendations on how the purposes, shape, structure, size and

funding of higher education, including support for students, should develop to meet the needs of the United Kingdom over the next 20 years" (NCIHE, 1997, 3). This would necessarily explore how to fund the continued expansion of the higher education system whilst maintaining or improving academic standards. Running parallel to the Dearing Committee, the JPG published its first report in April 1996. The report proposed the creation of a single body, independent of government and the funding councils, that would operate a largely unchanged *audit* and *assessment* approach (JPG, 1996b). The subsequent reaction from institutions was negative; it was felt that the key issue, the burden of the quality regime, had not been resolved. Furthermore, there was no substantive increase in self-evaluation.

The following year, after much deliberation between institutional representatives and the Department for Education, a plan for a single quality assurance body was agreed (JPG, 1996a). The dual *audit* and *assessment* approach would continue, however, it would be as part of a six-to-eight-year institutional quality assurance plan agreed between the agency and each institution. This would enable institutions greater control in the planning, timing and number of subject-level assessments within a national framework. Representatives from the provider under review would be able to observe these assessments, but would not be able to actively participate in them. These assessments would then form the main source of evidence for institution-wide peer reviews focusing on the management of quality (Brown, 2004, p.115-6).

The Quality Assurance Agency (QAA) was subsequently incorporated as a company on 27th March 1997, four months prior to the publication of the Dearing report, with a budget of 80% of the total for HEQC and the funding councils' quality assessment divisions. QAA officially took over the HEQC's staff and functions on 1st August 1997 and began operations, including its delegated responsibility for *assessment*, under contract with the national funding councils (Brown, 2004). Subscription to QAA was a requirement of receiving funds from the national funding councils. The new agency's board was made up of four members representing HEIs, four members representing the Funding Councils, and six independent members. Whilst owned by the representative bodies, the QAA was legally independent with a majority of its board members coming from elsewhere (QAA, 1997).

In summary, there were strong concerns over the burden, duplication and efficacy of the dual *audit* and *assessment* system. These concerns were exacerbated by the expansion of both approaches over time. In 1995, the Secretary of State responded and set out her criteria for a new unified system of quality assurance: it had to reduce burden, be cost-effective, maintain quality

and standards, and respect academic freedom. After two years of negotiation the Quality Assurance Agency was formed to take responsibility for both quality *audit* and *assessment* and do so by conducting cyclical reviews once every six to eight years.

2.3. The Evolution of the QAA

There have been three key phases in the development of the QAA prior to its current 'risk-based' incarnation. The QAA was established with the aim of reducing the burden on providers by combining the quality *audit* and *assessment* bodies. Whilst continuing the 'subject review' and institutional audit work of its predecessor bodies, the QAA sought to find a method that effectively combined its assessment and audit roles.

In July 1997, before the QAA could begin its first consultation, the comprehensive, 1,700 page Dearing report was published containing 93 executive recommendations. Beyond the headline recommendations of a flat-rate, non-means tested tuition fee and means-tested maintenance grants, there were a number of suggestions regarding quality that were well received by the Government (Stevens, 2005). The report stated the need for the standards of institutional awards to be maintained and recommended that the QAA provide benchmarking information and create a 'National Qualifications Framework' allowing it to "ensure that diversity is not an excuse for low standards or unacceptable quality" (NCIHE, 1997, p.143). The framework would detail what could be expected from each level of higher education award, regardless of the provider of the award, and a more detailed description of the skills and competencies associated with award holders. The QAA should then "work with universities and other degree-awarding institutions to create, within three years, a UK-wide pool of academic staff … from which institutions must select external examiners" (NCIHE, 1997, p.373) who would validate whether programmes met agreed standards for a level of award. Eight months later the QAA published its first consultation paper which broadly followed the suggestions of the 'Dearing Report': a pool of registered external examiners would check against a national standards framework as part of an approach which would allegedly have a 'lighter-touch' overall, either through less frequent reviews or a reduced review intensity when it does occur (QAA, 1998a).

Two years previously HEQC had consulted on a similar approach where external examiners, registered on a national database, would ensure the comparability of standards. The approach was abandoned, however, after a significant minority of universities objected on the grounds of the resources that would be required (HEQC, 1996). Five months later, following a second negative reaction to the suggestion of registered external examiners the idea was withdrawn; however, the

idea of benchmarking and programme specifications, equally unpopular with the sector (see, for example (Wolf, 1998)), was continued due to the positive reaction of students and employers to the consultation (QAA, 1998b).

A subsequent QAA consultation was launched in October 1998 and, following protracted and challenging negotiations between QAA, the Department for Education and Employment (DfEE), HEFCE and the representative bodies, an agreement resulting in a system more comprehensive than those it had replaced was reached for England. The new approach would entail 'Programme Reviews' conducted by academics that would result in a threshold judgement concerning programme outcome standards, i.e. were intended programme outcomes appropriate, were they achieved and if so were they maintainable? Furthermore, the 'Programme Reviews' would judge aspects of a provider's learning opportunities to be either 'commendable', 'approved' or 'failing'. In addition to 'Programme Reviews', 'Institutional Review' would examine the management of institutional standards and result in the identification of areas where it was essential, advisable or desirable for the institution to take action along with an overall judgement of confidence.

The new methodology, detailed in the QAA's *Handbook for Academic Review* (QAA, 2000), was far more comprehensive than those it had replaced. There were to be qualifications frameworks for every award, programme specifications for every course, benchmarks for every major subject, and codes of practice for each aspect of quality assurance (Brown, 2004). The approach was introduced in Scotland in October and was scheduled for implementation 12 months later in England; however, the process was to be short-lived and the QAA would soon enter its second phase.

The drive for closer attention to be paid to academic standards was led by the QAA's first Chief Executive, John Randall, who had vast ambitions for the inspection of higher education institutions and in 2000 wrote:

> "Subject benchmark information, programme specifications that spell out outcomes to be achieved, and a qualifications framework based on clear and explicit descriptors of level are the new means of defining standards in higher education. Together, they have a function similar to that of a code defining professional standards, in that they tell the individual client (the student) and the wider interested public (especially the employer) what they can reasonably expect from a professional service. Universities and their teachers must deliver to those standards if they are to convince the world that they are true professionals."

(Randall, 2000, p.166).

In 2001, fulfilling a recommendation of the Dearing Report and building on earlier work of HEQC, QAA published its 'Academic Infrastructure'. The set of reference points was developed "to help UK higher education providers to set, maintain and assure the academic standards of the higher education awards they make and the quality of the learning opportunities they offer" and ensure "broad comparability of standards at a threshold level across a diverse and dynamic sector" (QAA, 2010, p.2).

In the same year that QAA published its comprehensive 'Academic Infrastructure', however, the then Secretary of State, David Blunkett announced a 40% reduction in the volume of external review activity with departments that had scored highly in their last assessment becoming exempt from the next round. There were a number of factors that led to this announcement. In August 2000, a report commissioned by HEFCE put the annual cost to the sector of subject reviews at £30 million (PA Consulting Group, 2000). In January 2001, economists from the University of Warwick which had scored full marks in their recent subject review bitterly attacked the process (Harrison *et al.*, 2001). At the same time, Number 10 was being lobbied by a group of prominent Russell Group Vice-chancellors to reduce the quality assurance burden. Finally, the academic board of the LSE refused QAA reviewers access to the university and determined to "secede from [its] engagements with the QAA" which it believed had "infringed academic freedom, imposed its own bureaucratic and pedagogical agenda, neglected students' intellectual development and used incompetent and unprofessional reviewers" (THE, 2001b).

All these tensions were in evidence in the House of Lords on the evening following the Secretary of State's announcement when a debate was initiated by Lord Norton of Louth, a Professor at the University of Hull, and supported by other prominent peers on the complex and bureaucratic regulation of higher education (HL Deb, 2001). HEFCE, not the QAA, was subsequently tasked with leading the development of a new, less burdensome approach. Facing a reduction in subject-level assessment to an amount he considered unacceptable and, more importantly, wishing to take quality assurance in a different direction to HEFCE and UUK, John Randall resigned in August 2001 (Clare, 2001).

The following year, shortly before the QAA was to publish its revised approach, the Better Regulation Task Force, part of the 'New Labour' Government's drive for more targeted, less-burdensome regulation published *Higher Education: Easing the Burden* (BRTF, 2002). It too argued that higher education institutions were over-regulated, largely due to a lack of co-ordination between agencies, and recommended the strengthening of the Funding Council's HE Forum to

provide a "gatekeeper role to prevent unnecessary new burdens being placed on HEIs" (BRTF, 2002, 11).

Without universal application, 'Programme Review' was no longer an appropriate basis for a comprehensive regime. Having some programmes and institutions publicly judged against standards whilst the majority weren't would provide a piecemeal and confusing view of quality. The second key phase in the evolution of the QAA therefore began in 2002 with the proposal for a new approach of institutional-level review, based on *audit*, with subject-level *assessments* only undertaken on a highly selective basis where *audit* revealed concerns. In transitioning to the revised approach assessments would continue, covering up to a maximum of 10% of the institution's students. 'Principal' judgements were then made on:

- The level of confidence that can reasonably be placed in the soundness of the institution's management of the quality of its programmes and the academic standards of its awards; and, through direct scrutiny of primary evidence, whether the institution is securing acceptable academic standards and quality;
- The level of reliance that can reasonably be placed on the accuracy, integrity, completeness and frankness at the information that an institution publishes about the quality of its programmes and the standards of its awards.

(QAA, 2002)

The quality assurance of UK higher education was now, for all intents and purposes, entirely audit based. Every new course had to be validated by the provider, the provider would then monitor the course each year to ensure it was on track, and review it every five years or so against nationally agreed reference points known as 'Subject Benchmarks'. In addition, providers would continue the established system of appointing external examiners to check the quality and standards of each course. On a five-year cyclical basis, QAA were responsible for "checking these checks", i.e. that each provider had in place the necessary processes to ensure the quality and standards of their own provision meet expectations (QAA, 2012a). Teaching was not observed and quality outcomes were not assessed.

There was a price to be paid for the reduction in the quality burden however. A greater amount of quality information about each institution was now to be published, including external examiners' reports. Following a consultation, the QAA published its *Handbook for Institutional Audit* (2002) and the first of the new audits took place in February 2003. The Scottish took a slightly different approach and agreed to institutional audits less the 'discipline audit trail' assessments,

an internal subject review process and a separate quality enhancement process (QAA, 2003). Following a consultation, the *Handbook for Enhancement Led Institutional Review: Scotland* was published by the QAA in April 2003. The Welsh institutions and funding council also took the opportunity to reflect on the quality arrangements in England and opted for a revised, more enhancement-based approach without the requirement to publish external examiners' reports or internal reviews.

The Quality Assurance regime then remained relatively steady up until 2009 whilst publicly available information on higher education expanded rapidly. A fairly extensive set of information was made available on the 'Teaching Quality Information' website, now 'Unistats', in late 2004. The National Student Survey (NSS) was first run the following year and the results made available online. Throughout this time a number of high-profile quality failings occurred both in the UK and overseas, many highlighted by the QAA themselves. London Met was discovered to be offering modules in curry making and kite flying (THE, 2002). The University of Luton was found to have unsupervised recruitment agents in Bangladesh, Pakistan, India and Nigeria making admission decisions including waiving entry criteria (THE, 2005b). The University of Humberside was accused of having a branch in an Israeli petrol station and issuing 5,500 bogus degrees (MacKinnon and Norfolk, 2004). Leeds Met allowed "practical avoidance of challenging modules" (THE, 2005c). De Montfort increased grades in pharmacy by 14 per cent (THE, 2005a). These reports of quality failings culminated in a critical report by the House of Commons Innovation, Universities, Science and Skills Committee in 2009, which expressed concerns over low and declining academic standards in some areas of the UK higher education system and challenged the QAA's existing practices (IUSSC, 2009b).

The ongoing battle between self-regulation and academic freedom and an external agency prescribing standards was once again resumed with the committee making strong recommendations for change at the QAA. The committee only withheld calls for the replacement of the QAA on the grounds that "the inevitable hiatus, disruption and costs caused by the abolition of the QAA and establishment of a new body would not serve the best interests of students, universities and the taxpayer" (IUSSC, 2009b, p.97). The committee disagreed with the QAA's Chief Executive's, Peter Williams', view that process and outcomes "were very strongly linked" (IUSSC, 2009b, p.94)and stated that "in not judging the standards themselves, the QAA is taking an unduly limited view of its potential role" (IUSSC, 2009b, p.97). The committee recommended that the QAA should be reformed and re-established as a Quality and Standards Agency which:

- Had a duty to safeguard, and report on, standards in higher education in England.

- Should be half funded through HEFCE and half from levies on HEIs in England to ensure its independence.
- Should review and report on the quality of teaching in universities and, where shortcomings are identified, ensure that they are reported publicly and addressed by the institutions concerned.
- Have powers to carry out reviews of the quality of, and standards applied in, the assessment for an institution's courses, including, if necessary, its degree awarding powers, in response to external examiners' or public concerns about the standards in an institution or at the direction of the Secretary of State.

(IUSSC, 2009b)

The IUSSC, therefore, were calling for the QAA to take a tougher stance on the variation in academic standards and focus on direct assessment of teaching quality rather than the processes that facilitated it. Were fundamental changes to the operation of the QAA not achieved within two years, the IUSSC recommended that the "QAA/Quality and Standards Agency should be abolished and an entirely new organisation be established in its place" (IUSSC, 2009b, p.97).

The demands of the IUSSC heralded the third key phase in the life of the QAA. A number of changes were made and in 2011/12, the 'Institutional Audit' methodology was replaced by the 'Institutional Review' (QAA, 2011a). Further to *academic standards* and *the quality of teaching and learning*, *the provision of information* and *enhancement* would be scrutinised by reviewers, the resulting judgements would fall into four, not three, categories, and students would become full members of review teams (Brown and Carasso, 2013). The revised approach necessitated a review of QAA's 'Academic Infrastructure' which was renamed the 'UK Quality Code for Higher Education' (the 'Quality Code') and expanded to include additional reference points, or 'expectations', for the expanded review method (QAA, 2011b).

In summary, with the creation of a unified quality assurance system looking set to produce something more comprehensive than the dual *audit* and *assessment* system it had replaced, the Secretary of State intervened and called for a reduction in regulatory burdens. What resulted was an almost entirely *audit* based approach in which institutional-level quality assurance processes were checked for every provider on a cyclical basis and outcomes were not assessed. Whilst less burdensome, the new approach failed to stem ongoing concerns over falling quality and standards

in an expanding sector and 'institutional audit' was superseded by the more comprehensive, wide-ranging 'institutional review'.

2.4. A Risk-Based Approach to Quality Assurance

Shortly after the 'Institutional Review' method was finalised the Government published the 2011 White Paper *Students at the Heart of the System.* Having already legislated to raise the cap for undergraduate tuition fees from £3,000 to £9,000 to shift the cost of higher education to the 'consumer' and promote price competition in the sector, the White Paper contained further proposals to marketise the sector. Students at new, for-profit 'alternative providers' would be eligible for funding from the Student Loans Company, whilst the providers themselves required fewer students to obtain the 'university' title and would find it easier to obtain their own degree-awarding powers. All providers faced further requirements for the provision of information to support 'consumer' choice.

A combination of new, inexperienced providers entering the market – many with profit-driven motives – and increased competition amongst existing providers led to a change of emphasis for the QAA. *Students at the Heart of the System* proposed:

> "…a genuinely risk-based approach, focusing QAA effort where it will have most impact and giving students power to hold universities to account. All providers must continue to be part of a single assurance framework. But we would explore options in which the frequency – and perhaps need – for a full, scheduled institutional review will depend on an objective assessment of a basket of data, monitored continually but at arm's length."
>
> (BIS, 2011, 3.19)

Now, rather than all providers being reviewed at the same frequency regardless of their size, type, or performance, data were to be used to direct QAA's activity, prioritising those providers at which the QAA was most likely to identify quality assurance issues. The White Paper asked HEFCE to consult on the criteria for assessing risk and the resulting frequency and intensity of reviews "with a view to achieving very substantial deregulatory change for institutions that can demonstrate low risk" (BIS, 2011, 3.20). *Ad hoc* triggers which could prompt an otherwise unscheduled review by the QAA would also be subject to consultation.

The subsequent consultation revealed significant concerns over the "scope, validity, availability and reliability" of metrics that might form part of a risk-based quality assurance regime. Plans for

QAA to regularly monitor a basket of data were not continued (HEFCE, 2012b). Following the consultation, Alan Langlands, the Chief Executive of HEFCE, wrote to Anthony McClaren, Chief Executive of QAA, inviting the QAA "to implement a more risk-based approach to the quality assurance of higher education in England" with a view to implementing the revised approach during the 2013/14 academic year (Langlands, 2012, p.1). The result was the revised 'Higher Education Review' (HER) approach for HEIs and FECs, designed to target the QAA's efforts where they are most needed and tailor external reviews to the individual circumstances of providers being reviewed (QAA, 2013a).

The revised 'Higher Education Review' (HER) approach was 'risk-based' in only the most limited of ways. 'Risk-based' approaches are designed to prioritise regulatory resource to areas where the risk, defined as the product of the impact and the likelihood of a regulatee not meeting regulatory standards, is greatest (Rothstein *et al.*, 2006b). Under the *HER* approach, institutions whose two previous institution-wide reviews were regarded as successful would continue to be reviewed every six years. Institutions that were either not successful in their last two reviews, had not yet been reviewed twice, had concerns upheld about the quality of their provision following a full inquiry under the QAA's concern scheme, or had undergone significant managerial change would be reviewed four years after their previous review (QAA, 2013a). The 'impact' of different providers not meeting the standards detailed in the *Quality Code* was not considered.

The new 'alternative providers' could be subject to QAA review for two reasons. *First*, for alternative providers looking for their students to have access to funding from the Student Loan Company (SLC), QAA operated the 'Review for Specific Course Designation' (RSCD) programme. Only once a course has been approved, or 'designated', by QAA were students able to use government loans to pay for their tuition fees and maintenance (QAA, 2014b). *Second,* alternative providers seeking to recruit non-EU students directly must undergo a QAA 'review of educational oversight' (REO) as part of their application for, and continued ownership of, 'Tier 4 Sponsor Status'[1] (QAA, 2012c). Both the RSCD and REO processes comprised a Financial Sustainability, Management and Governance (FSMG) check conducted by a third party, and a QAA review of all higher education provision addressing the same four key areas as at universities and colleges: *academic standards*, *the quality of teaching and learning*, *the provision of information*, and *enhancement*. Following a successful RSCD or REO application, alternative providers were

[1] Holding 'Tier 4 Sponsor Status', previously known as 'Highly Trusted Status', allows education providers to sponsor international students to come to the UK under Tier 4 of the UK's points based visas and immigration system. That is, education providers are trusted with the responsibility to approve visas for their own foreign students coming to study with them.

required to complete an annual information return. Should any material changes or concerns be identified then a review would take place outside the normal review cycle.

There were, and continue to be, significant concerns about the quality of a number of new, alternative providers. In 2015, the Public Accounts Committee (PAC) reported very high drop-out and absence rates, poor administration and inappropriate recruitment practices, including colleges recruiting on the streets and students being accepted onto courses while lacking adequate English language skills (PAC, 2015). The report did not criticise QAA's approach, but did express further concern over the rapid expansion of alternative providers who, unlike universities and colleges, initially faced no cap on the number of students they could recruit. For example, by the time the Government introduced controls on alternative providers in November 2013, having spent far more than it had budgeted for, Regent's College had expanded from having 10 Higher National Diploma (HND) students to over 1,000. Similarly, the intake at St. Patrick's College ballooned from 50 to over 4,000 in one year (McGettigan, 2014). Although the cap on student numbers was lifted for universities and colleges for the 2015/16 academic year, the cap remains in place for alternative providers (Hillman, 2014; Clark, 2015).

Meanwhile, in October 2014, to the surprise of many, HEFCE, HEFCW and DELNI (the funding bodies for England, Wales and Northern Ireland respectively, though not Scotland) announced that they were to "seek views on future approaches to the assessment of quality in higher education" and based on the feedback received "design a specification and invite tenders under a joint procurement exercise" (Atkins, 2015). The funding councils' responsibility to ensure the quality of the provision they fund established in the 1992 *Further and Higher Education Act* had, up until this point, been contracted out to the QAA without a competitive tender. The review, purportedly conducted to "ensure transparency and demonstrate value for money" (Atkins, 2015) was seen by many as a power grab by the funding councils whose roles allocating funding had been greatly diminished by the changes to tuition fees and lifting of the English student numbers cap (see for example: Million+, 2014; WonkHE, 2015). The funding councils were looking to develop "innovative approaches that are risk-based, proportionate, affordable, and low burden" suitable for the "fast-evolving and increasingly diverse higher education environment" (Atkins, 2015).

The *Quality Assessment Review Steering Group* (QARSG) formed by HEFCE, HEFCW and DELNI held a two-part consultation. In January 2015 a 'discussion document' was published to "explore the deep, critical questions that need to be addressed before the more practical issues surrounding the design and implementation of any new quality assessment arrangements can be considered"

(QARSG, 2015c). The document laid out the vision of a changing higher education sector: expanding, more global, with new 'alternative' and online providers, and greater student expectations. Amongst other things the QARSG sought agreement on the principle that the future quality assessment arrangement should be risk-based and asked "what evidence and/or data should be used to identify quality issues in an individual provider?" (QARSG, 2015c). The majority of stakeholders felt that a risk-based approach should be adopted, focusing attention on less-established providers and those whose performance measures suggests they are, or may in the future be, a cause for concern. Suggested performance measures included: provider type, mission and age, staff turnover, student numbers, retention, satisfaction, and earnings data (MRUK Research, 2015).

The future of the quality assurance of higher education became yet more uncertain in the spring of 2015, however, when the Conservatives won an outright majority on a manifesto containing promises of a 'Teaching Excellence Framework', and the QARSG published phase two of their consultation (QARSG, 2015b; Conservative Party, 2015). The QARSG's plans for a new quality assessment regime and the Government's plans for the TEF appeared to have been developed in isolation. How the two would co-exist was not clear.

QARSG's phase two consultation again proposed a general risk-based approach along with additional changes including: a switch from focusing on processes to outcomes, an enhanced role for university's governing bodies in assuring quality, and the oft proposed and rejected idea of a register of external examiners. The results of the consultation were published in November 2015 and all proposals, except for making governors responsible for quality assurance and the training and registration of external examiners, were accepted by the majority of respondees (QARSG, 2015a).

Also in November 2015, the Government published its Green Paper *Fulfilling Our Potential: Teaching Excellence, Social Mobility and Student Choice* (2015b) which contained three key proposals that would impact quality assurance. *First,* the 'regulatory playing field' was to be levelled, again making it easier for new providers to enter the market and to obtain the 'university' title and degree awarding powers. *Second,* a Teaching Excellence Framework (TEF), as part of an enhanced higher education market, would force providers to improve teaching quality or drive them out of the sector. TEF judgements would be made by panels of experts and informed by existing, benchmarked metrics of graduate employment, retention rates, and student satisfaction. *Third,* a number of regulators were to be merged to form the 'Office for Students' (OfS) bringing together access agreements, quality assurance, data collection and processing, award of the

'university' title, TEF and teaching funding into a single student champion organisation (BIS, 2015b).

As the consultation on the Green Paper took place, HEFCE, HEFCW and DELNI published their revised operating model for quality assessment (HEFCE, 2016). The operating model confirmed that future quality assurance work would be conducted by themselves and the winners of six competitive tenders. Four of the six contracts, including the largest contract relating to conducting quality assurance reviews, were awarded to QAA (QAA, 2016a).

Under the revised process established providers are to no longer face cyclical reviews. Instead, experts at the national funding councils will conduct a desk-based 'Annual Provider Review' exercise that will "build on established data analysis and assurance arrangements" to determine those providers in need of a quality assurance review (HEFCE, 2016, 92). The 'Annual Provider Review' will consider metrics on *inter alia* continuation rates, degree outcomes, student satisfaction, and employment outcomes. New providers are to enter a four-year probationary period which will include an initial peer review. During the four years of 'development', new providers will be subject to the same desk-based 'Annual Provider Review' as 'established' providers. After four years, probationary providers will undergo a peer review and either be judged 'established', or remain in the probationary process until such time as they become 'established' (HEFCE, 2016c).

How long this system will last is unclear. In May 2016 the Government published its White Paper that retained the three key quality assurance measures of the Green Paper: the regulatory landscape is to be 'levelled' and will move towards a risk-based approach, the TEF is being trialled in 2016/17, and a number of regulators are to be merged to form the 'Office for Students' (OfS). Somewhat pointedly, the White Paper stated:

> "the Secretary of State will have a power to designate sector-owned organisations to carry out specific quality assurance and data publication functions, enabling the principle of co-regulation to continue."
>
> (BIS, 2016, p.18)

leading to suggestions that, at the Government's wishes, all quality assurance work will be returned to QAA (WonkHE, 2016). As of late November 2016, the higher education bill derived from the White Paper has passed its third reading stage and is progressing to the House of Lords.

In summary, the higher education landscape has changed dramatically since 2011. A stable collection of established HEIs and FECs have been joined by new 'alternative providers'. At the same time tuition fees have been tripled and a host of measures have been enacted to create more of a higher education market with the intention that the market will drive up quality. With more providers entering the market the QAA's approach became 'risk-based', albeit to a very limited extent with high-risk providers only being reviewed four, rather than six, years after their previous review and without the focus on "a basket of data" envisioned in *Students at the Heart of the System* (BIS, 2011, 3.19). A new metric-led, and outcomes focused, risk-based approach is now being introduced however following the *Quality Assessment Review*. This approach may yet change again if and when the higher education bill receives full parliamentary approval.

2.5. Conclusion

Quality Assurance agencies have developed over the past 30 years for a number of reasons. *First*, a continuous 'marketisation' of higher education, seen as the most efficient means of achieving high standards, and an increasing reliance on institutions being able to attract students to survive, has introduced increased risks to quality including low standards and grade inflation. *Second*, the number of students has increased dramatically; whilst one in seven further education leavers entered higher education in 1987, nearly one in two entered higher education in 2011. With the increase in student numbers has come an increase in public expenditure, £15 billion in 2009/10 (IUSSC, 2009a), and with it the need for increased accountability. *Third*, since universities were brought under central control in the late 1980s, the political fortunes of the party in power have become aligned to the performance of the higher education sector. Politicians have a more direct incentive to ensure quality is being maintained or improved.

The QAA itself was deemed necessary as the work of its predecessors, HEQC and the Quality Assessment Councils, was perceived as overlapping and too burdensome. The QAA has since been through a number of phases itself suffering from the competing demands of 'low burden' oversight and respecting academic freedom on the one hand, and comprehensive standard-setting and assessment on the other. Initially it struggled to fulfil its remit of combining audit and assessment effectively whilst reducing the burden imposed on higher education institutions. The Secretary of State subsequently demanded a 40 per cent reduction in burden. A 'lighter touch' approach focusing on the institution-level audit of providers' processes was introduced and teaching quality was no longer directly assessed. The pendulum then proceeded to swing back

towards more comprehensive oversight following a critical IUSSC report (2009b) which deemed the QAA not to be doing enough to define and enforce standards.

Higher education began a series of significant changes in 2011. Higher education providers were no longer the same established collection of HEIs and FECs but were being joined by 'alternative providers', some with little experience and the goal of shareholder returns. The introduction of new providers, along with the additional pressures of increased competition for existing providers, resulted in the QAA being asked to adopt a metric-led, risk-based approach to prioritising their review activity. Following consultation, the adopted HER, RSCD and REO approaches were very limited in terms of being 'risk-based' and data driven. As the sector continues to evolve HEFCE has recently implemented a new metric-led, risk-based approach to the quality assurance of higher education (HEFCE, 2016c). This may yet change again if the higher education bill currently working its way through parliament is approved. A more in-depth assessment of what constitutes a 'risk-based' approach, its merits and limitations, and suggested applications to quality assurance in higher education are explored in the next chapter.

Appendix A – Summary of Major Events

1983 The Secretary of State, Sir Keith Joseph, writes to the University Grants Committee (UGC) asking them assess how standards were currently being maintained and enhanced and, in the context of a more efficient use of resources in universities, to explore the possibilities for maintaining and improving academic quality in the future.

1986 The report of the 'Reynolds Group', established by the CVCP to study and report on institutions' methods for maintaining and monitoring academic quality and standards in response to the 1983 request by the Secretary of State, recommends points of reference for the self-assurance of the maintenance and monitoring of standards.

1988 The 1988 White Paper establishes the Polytechnics and Colleges Funding Council (PCFC) and the Universities Funding Council (UFC) who would no longer fund institutions through grants but rather would operate a system of contracting. Greater control could now be exerted over higher education institutions.

1990 The Academic Audit Unit (AAU), designed to provide a more permanent oversight of universities' standards and quality control, commences operations.

1992 The binary divide ends with 'public sector' institutions, under certain circumstances, able to adopt the title of 'university' and the PCFC and UFC being replaced by national funding councils.

Quality Assessment Councils are established within the funding councils and undertake assessments of the performance of subject areas within universities against their own declared aims and objectives.

HEQC is founded and begins its quality audits where three peer auditors visit an institution to focus on four main areas: the provision and design of course and degree programmes, teaching and communication methods, academic staff, and means of taking account of external examiners' reports and the views of students' and external bodies'.

1994 The Secretary of State, John Patten MP, gives a speech at the HEFCE annual conference stressing the importance the Government placed upon the broad comparability of standards.

Assessment visits are extended to cover all institutions and departments and judgements of 'excellent', 'satisfactory' or 'unsatisfactory' are replaced by a score of 1 (the lowest) to 4 (the highest) for each of six core areas leading to a published overall profile of judgements containing a score out of 24.

The new Secretary of State, Mrs Gillian Shephard, asks the HEFCE Chief Executive, Professor (now sir) Graeme Davies, to work with the representative bodies to determine how audit and assessment could be combined to form a single quality system.

1995 A Joint Planning Group is established by HEFCE and the CVCP to develop proposals for a single quality regime.

1996 The National Committee of Inquiry into Higher Education or the "Dearing Committee" is established by the Secretary of State to "make recommendations on how the purposes, shape, structure, size and funding of higher education, including support for students, should develop to meet the needs of the United Kingdom over the next 20 years". This explores how to fund the continued expansion of the higher education system whilst maintaining or improving academic standards

1997 The Quality Assurance Agency is incorporated and takes over the HEQC's staff and functions. The new agency's board was made up of four members representing HEIs, four members representing the Funding Councils, and six independent members. Representing a shift away from pure self-regulation and towards a partnership between the state and the institutions.

The comprehensive Dearing report is published containing 93 executive recommendations. The report suggests the need for the standards of institutional awards to be maintained and recommended and that the QAA should provide benchmarking information, create a 'National Qualifications Framework' and create a UK-wide pool of external examiners.

1998 The QAA undertakes its first and second consultations and, following negotiations between the Department for Education and Employment (DfEE), HEFCE, representative bodies and QAA, an agreement resulting in a system more comprehensive than those it had replaced was reached for England.

The new approach entails 'Programme Reviews', resulting in threshold judgements concerning programme outcome standards and graded judgements for a number of aspects of learning opportunities, and 'Institutional Reviews' resulting in an overall judgement of confidence in an institution's management of standards.

2000 The QAA publishes the *Handbook for Academic Review* detailing its finalised methodology.

2001 LSE's Academic Board make a stand and threaten to withdraw from the QAA over concerns about the new approach's infringement of academic freedom and the low standard of reviewers.

The Secretary of State, David Blunkett, announces a 40% reduction in the volume of external review activity with departments that have scored highly in their last assessment becoming exempt from the next round.

John Randall, Chief Executive of the QAA, resigns.

2002 The QAA publishes its *Handbook for Institutional Audit* detailing the revised, lighter-touch approach following the Secretary of States announcement the previous year.

2004 A fairly extensive set of information is made available on the *Teaching Quality Information* website, now *Unistats*.

2005 The first National Student Survey (NSS) is run.

2009 A critical report by the House of Commons Innovation, Universities, Science and Skills Committee expresses concerns over low and declining academic standards in some areas of the UK higher education system and challenges the QAA's existing practices stating that "in not judging the standards themselves, the QAA is taking an unduly limited view of its potential role" (IUSSC, 2009, p.97). The committee recommends that the QAA was to be reformed and re-established as a Quality and Standards Agency.

2010 Parliament votes to raise the maximum tuition fees for undergraduate degrees provided by centrally-funded institutions to £9,000 per annum.

2011 The 'Institutional Audit' methodology is replaced by the 'Institutional Review (England and Northern Ireland)' seeing four areas now scrutinised by reviewers instead of two, the resulting judgements falling into four, not three, categories, students becoming full members of review teams, and there no longer being fixed programme of reviews but instead a more responsive rolling programme.

BIS publishes the White Paper *Students at the Heart of the System* proposing a risk-based approach for the QAA, focusing its effort were it will have greatest impact. New providers may receive a more regular and in-depth review whilst those providers with a demonstrable track record of high-quality provision can expect to see significantly less frequent or in-depth reviews.

2012 A HEFCE consultation reveals significant concerns over the "scope, validity, availability and reliability" of metrics which might form part of a risk-based quality assurance regime and plans for QAA to regularly monitor a basket of data are not continued.

2013 QAA introduces its 'risk-based' HER approach for HEIs and FECs.

2014 HEFCE, HEFCW and DELNI form the *Quality Assessment Review Steering Group* and announce that they are to "seek views on future approaches to the assessment of quality in higher education" and based on the feedback received "design a specification and invite tenders under a joint procurement exercise".

2015 QARSG's consultation finds support for a risk-based approach to quality assurance focusing on processes rather than outcomes.

The Conservative Party wins an outright majority and a Higher Education Green Paper is published proposing a 'Teaching Excellence Framework'.

The BIS Select Committee announces an inquiry into 'Assuring Quality in Higher Education'.

2016 HEFCE puts quality assurance work out to tender with QAA winning four of the six available contracts.

The revised quality assurance regime sees an end to cyclical reviews and the introduction of a metric-led, risk-based approach.

The higher education bill is presented to parliament maintaining the key proposals laid out in the Green and White Papers.

Appendix B - The UK Quality Code for Higher Education

The Quality Code (2011) is made up of three parts. Part A *Setting and Maintaining Academic Standards* contains six of the 19 'expectations'. These six expectations are not accompanied by any 'indicators of sound practice'. Part B *Assuring and Enhancing Academic Quality* contains 11 of the 19 'expectations' each accompanied by multiple indicators. Part C *Information about Higher Education Provision* contains a single 'expectation' and five accompanying 'indicators of sound practice'. Each 'expectation' and its accompanying 'indicators of sound practice' is detailed in turn below.

Part A Setting and Maintaining Academic Standards

Expectation A1

In order to secure threshold academic standards, degree-awarding bodies:

a) ensure that the requirements of The framework for higher education qualifications in England, Wales and Northern Ireland/The framework for qualifications of higher education institutions in Scotland are met by:
   - positioning their qualifications at the appropriate level of the relevant framework for higher education qualifications
   - ensuring that programme learning outcomes align with the relevant qualification descriptor in the relevant framework for higher education qualifications
   - naming qualifications in accordance with the titling conventions specified in the frameworks for higher education qualifications
   - awarding qualifications to mark the achievement of positively defined programme learning outcomes

b) consider and take account of QAA's guidance on qualification characteristics

c) where they award UK credit, assign credit values and design programmes that align with the specifications of the relevant national credit framework.

d) consider and take account of relevant subject benchmark statements.

Expectation A2.1

In order to secure their academic standards, degree-awarding bodies establish transparent and comprehensive academic frameworks and regulations to govern how they award academic credit and qualifications.

Expectation A2.2

Degree-awarding bodies maintain a definitive record of each programme and qualification that they approve (and of subsequent changes to it) which constitutes the reference point for delivery and assessment of the programme, its monitoring and review, and for the provision of records of study to students and alumni.

Expectation A3.1

Degree-awarding bodies establish and consistently implement processes for the approval of taught programmes and research degrees that ensure that academic standards are set at a level which meets the UK threshold standard for the qualification and are in accordance with their own academic frameworks and regulations.

Expectation A3.2

Degree-awarding bodies ensure that credit and qualifications are awarded only where:

- the achievement of relevant learning outcomes (module learning outcomes in the case of credit, and programme outcomes in the case of qualifications) has been demonstrated through assessment
- both the UK threshold standards and the academic standards of the relevant degree-awarding body have been satisfied.

Expectation A3.3

Degree-awarding bodies ensure that processes for the monitoring and review of programmes are implemented which explicitly address whether the UK threshold academic standards are achieved and whether the academic standards required by the individual degree-awarding body are being maintained.

Expectation A3.4

In order to be transparent and publicly accountable, degree-awarding bodies use external and independent expertise at key stages of setting and maintaining academic standards to advise on whether:

- UK threshold academic standards are set, delivered and achieved
- the academic standards of the degree-awarding body are appropriately set and maintained.

Part B: Assuring and Enhancing Academic Quality

Expectation B1 (Programme design, development and approval)

Higher education providers, in discharging their responsibilities for setting and maintaining academic standards and assuring and enhancing the quality of learning opportunities, operate effective processes for the design, development and approval of programmes.

Indicators

1. Higher education providers maintain strategic oversight of the processes for, and outcomes of, programme design, development and approval, to ensure processes are applied systematically and operated consistently.
2. Higher education providers make clear the criteria against which programme proposals are assessed in the programme approval process.
3. Higher education providers define processes, roles and responsibilities for programme design, development and approval and communicate them to those involved.
4. Higher education providers evaluate their processes for programme design, development and approval and take action to improve them where necessary.
5. Higher education providers make use of reference points and expertise from outside the programme in programme design and in their processes for programme development and approval.
6. Higher education providers involve students in programme design and in processes for programme development and approval.
7. Higher education providers enable staff and other participants to contribute effectively to programme design, development and approval by putting in place appropriate arrangements for their support and development.

Expectation B2 (Recruitment, selection and admission)

Recruitment, selection, and admission policies and procedures adhere to the principles of fair admission. They are transparent, reliable, valid, inclusive and underpinned by appropriate organisational structures and processes. They support higher education providers in the selection of students who are able to complete their programme.

Indicators

1. Recruitment, selection, and admission policies are informed by the strategic priorities of the higher education provider. Higher education providers promote a shared understanding of their approach among all those involved in recruitment, selection, and admission.

2. Recruitment, selection and admission processes are conducted in a professional manner by authorised and competent representatives of the higher education provider.
3. Higher education providers have procedures for handling appeals and complaints about recruitment, selection and admission that are fair and accessible. Appeals and complaints procedures are conducted expeditiously and in accordance with a published timescale.
4. Higher education providers monitor, review and update their recruitment, selection and admission policies and procedures, in order to enhance them and to ensure that they continue to support the provider's mission and strategic objectives. Higher education providers determine the frequency with which monitoring and review are undertaken.
5. Recruitment activities undertaken by higher education providers assist prospective students in making informed decisions about higher education.
6. Higher education providers make clear to prospective students how the recruitment, selection and admission process will be conducted and what prospective students have to do.
7. Selection processes for entry into higher education are underpinned by transparent entry requirements, both academic and non-academic, and present no unnecessary barriers to prospective students.
8. Higher education providers determine how decisions and the reasons for those decisions are recorded and conveyed to prospective students.
9. Higher education providers inform prospective students, at the earliest opportunity, of any significant changes to a programme to which they have applied. Prospective students are advised promptly of the options available in the circumstances.
10. Higher education providers give successful applicants sufficient information to enable them to make the transition from prospective student to current student

Expectation B3 (Learning and teaching)

Higher education providers, working with their staff, students and other stakeholders, articulate and systematically review and enhance the provision of learning opportunities and teaching practices, so that every student is enabled to develop as an independent learner, study their chosen subject(s) in depth and enhance their capacity for analytical, critical and creative thinking.

Indicators

1. Higher education providers articulate and implement a strategic approach to learning and teaching and promote a shared understanding of this approach among their staff, students and other stakeholders.

2. Learning and teaching activities and associated resources provide every student with an equal and effective opportunity to achieve the intended learning outcomes.
3. Learning and teaching practices are informed by reflection, evaluation of professional practice, and subject-specific and educational scholarship.
4. Higher education providers assure themselves that everyone involved in teaching or supporting student learning is appropriately qualified, supported and developed.
5. Higher education providers collect and analyse appropriate information to ensure the continued effectiveness of their strategic approach to, and the enhancement of, learning opportunities and teaching practices.
6. Higher education providers maintain physical, virtual and social learning environments that are safe, accessible and reliable for every student, promoting dignity, courtesy and respect in their use.
7. Every student is provided with clear and current information that specifies the learning opportunities and support available to them.
8. Higher education providers take deliberate steps to assist every student to understand their responsibility to engage with the learning opportunities provided and shape their learning experience.
9. Every student is enabled to monitor their progress and further their academic development through the provision of regular opportunities to reflect on feedback and engage in dialogue with staff.

Expectation B4 (Enabling student development and achievement)

Higher education providers have in place, monitor and evaluate arrangements and resources which enable students to develop their academic, personal and professional potential.

Indicators

1. Through strategic and operational planning, and quality assurance and enhancement, higher education providers determine and evaluate how they enable student development and achievement.
2. Higher education providers define, coordinate, monitor and evaluate roles and responsibilities for enabling student development and achievement both internally and in cooperation with other organisations.
3. A commitment to equity guides higher education providers in enabling student development and achievement.
4. Higher education providers inform students before and during their period of study of opportunities designed to enable their development and achievement.

5. To enable student development and achievement, higher education providers put in place policies, practices and systems that facilitate successful transitions and academic progression.
6. Higher education providers ensure all students have opportunities to develop skills that enable their academic, personal and professional progression.
7. Higher education providers ensure staff who enable students to develop and achieve are appropriately qualified, competent, up to date and supported.
8. Higher education providers make available appropriate learning resources and enable students to develop the skills to use them.

Expectation B5 (Student engagement)

Higher education providers take deliberate steps to engage all students, individually and collectively, as partners in the assurance and enhancement of their educational experience.

Indicators

1. Higher education providers, in partnership with their student body, define and promote the range of opportunities for any student to engage in educational enhancement and quality assurance.
2. Higher education providers create and maintain an environment within which students and staff engage in discussions that aim to bring about demonstrable enhancement of the educational experience.
3. Arrangements exist for the effective representation of the collective student voice at all organisational levels, and these arrangements provide opportunities for all students to be heard.
4. Higher education providers ensure that student representatives and staff have access to training and ongoing support to equip them to fulfil their roles in educational enhancement and quality assurance effectively.
5. Students and staff engage in evidence-based discussions based on the mutual sharing of information.
6. Staff and students to disseminate and jointly recognise the enhancements made to the student educational experience, and the efforts of students in achieving these successes.
7. The effectiveness of student engagement is monitored and reviewed at least annually, using pre-defined key performance indicators, and policies and processes enhanced where required.

Expectation B6 (Assessment and the recognition of prior learning)

Higher education providers operate equitable, valid and reliable processes of assessment, including for the recognition of prior learning, which enable every student to demonstrate the extent to which they have achieved the intended learning outcomes for the credit or qualification being sought.

Indicators

1. Higher education providers operate effective policies, regulations and processes which ensure that the academic standard for each award of credit or a qualification is rigorously set and maintained at the appropriate level, and that student performance is equitably judged against this standard.
2. Assessment policies, regulations and processes, including those for the recognition of prior learning, are explicit, transparent and accessible to all intended audiences.
3. Those who might be eligible for the recognition of prior learning are made aware of the opportunities available, and are supported throughout the process of application and assessment for recognition.
4. Higher education providers assure themselves that everyone involved in the assessment of student work, including prior learning, and associated assessment processes is competent to undertake their roles and responsibilities.
5. Assessment and feedback practices are informed by reflection, consideration of professional practice, and subject-specific and educational scholarship.
6. Staff and students engage in dialogue to promote a shared understanding of the basis on which academic judgements are made.
7. Students are provided with opportunities to develop an understanding of, and the necessary skills to demonstrate, good academic practice.
8. The volume, timing and nature of assessment enable students to demonstrate the extent to which they have achieved the intended learning outcomes.
9. Feedback on assessment is timely, constructive and developmental.
10. Through inclusive design wherever possible, and through individual reasonable adjustments wherever required, assessment tasks provide every student with an equal opportunity to demonstrate their achievement.
11. Assessment is carried out securely.
12. Degree-awarding bodies assure themselves that the standards of their awards are not compromised as a result of conducting assessment in a language other than English.

13. Processes for marking assessments and for moderating marks are clearly articulated and consistently operated by those involved in the assessment process.
14. Higher education providers operate processes for preventing, identifying, investigating and responding to unacceptable academic practice.
15. Degree-awarding bodies specify clearly the membership, procedures, powers and accountability of examination boards and assessment panels, including those dealing with the recognition of prior learning; this information is available to all members of such boards.
16. Boards of examiners/assessment panels apply fairly and consistently regulations for progression within, and transfer between, programmes and for the award of credits and qualifications.
17. The decisions of examination boards and assessment panels are recorded accurately, and communicated to students promptly and in accordance with stated timescales.
18. Degree-awarding bodies systematically evaluate and enhance their assessment policies, regulations and processes.

Expectation B7 (External examining)

Higher education providers make scrupulous use of external examiners.

Indicators

1. Degree-awarding bodies appoint one or more external examiner(s) to carry out the role(s) defined in this Chapter for all provision that leads to a higher education qualification of the degree-awarding body.
2. Degree-awarding bodies expect their external examiners to provide informative comment and recommendations upon whether or not:
    - the degree-awarding body is maintaining the threshold academic standards set for its awards in accordance with the frameworks for higher education qualifications and applicable Subject Benchmark Statements
    - the assessment process measures student achievement rigorously and fairly against the intended outcomes of the programme(s) and is conducted in line with the degree awarding body's policies and regulations
    - the academic standards and the achievements of students are comparable with those in other UK degree-awarding bodies of which the external examiners have experience.
3. Degree-awarding bodies expect their external examiners to provide informative comment and recommendations on:

- good practice and innovation relating to learning, teaching and assessment observed by the external examiners
- opportunities to enhance the quality of the learning opportunities provided to students.

4. Degree-awarding bodies have explicit policies and regulations governing the nomination and appointment of external examiners. Degree-awarding bodies can terminate an external examiner's appointment at any time, subject to approved institutional procedures, for failure by the external examiner to fulfil his/her obligations or if a conflict of interest arises which cannot be satisfactorily resolved.
5. Degree awarding bodies apply the following UK-wide set of criteria for appointing external examiners and make every effort to ensure that their external examiners are competent to undertake the responsibilities defined by the institution. Degree awarding bodies use the criteria to ensure that potential conflicts of interest are identified and resolved prior to appointing external examiners or as soon as they arise.

Person specification

a) Degree-awarding bodies appoint external examiners who can show appropriate evidence of the following:
   i. knowledge and understanding of UK sector agreed reference points for the maintenance of academic standards and assurance and enhancement of quality
   ii. competence and experience in the fields covered by the programme of study, or parts thereof
   iii. relevant academic and/or professional qualifications to at least the level of the qualification being externally examined, and/or extensive practitioner experience where appropriate
   iv. competence and experience relating to designing and operating a variety of assessment tasks appropriate to the subject and operating assessment procedures
   v. sufficient standing, credibility and breadth of experience within the discipline to be able to command the respect of academic peers and, where appropriate, professional peers
   vi. familiarity with the standard to be expected of students to achieve the award that is to be assessed

vii. fluency in English, and where programmes are delivered and assessed in languages other than English, fluency in the relevant language(s) (unless other secure arrangements are in place to ensure that external examiners are provided with the information to make their judgements)

viii. meeting applicable criteria set by professional, statutory or regulatory bodies

ix. awareness of current developments in the design and delivery of relevant curricula

x. competence and experience relating to the enhancement of the student learning experience.

Conflicts of interest

b) b. Degree-awarding bodies do not appoint as external examiners anyone in the following categories or circumstances:

i. a member of a governing body or committee of the appointing body or one of its partners, delivery organisations or support providers, or a current employee of the appointing body or one of its partners, delivery organisations or support providers

ii. anyone with a close professional, contractual or personal relationship with a member of staff or student involved with the programme of study

iii. anyone required to assess colleagues who are recruited as students to the programme of study

iv. anyone who is, or knows they will be, in a position to influence significantly the future of students on the programme of study

v. anyone significantly involved in recent or current substantive collaborative research activities with a member of staff closely involved in the delivery, management or assessment of the programme(s) or modules in question

vi. former staff or students of the higher education provider unless a period of five years has elapsed and all students taught by or with the external examiner have completed their programme(s)

vii. a reciprocal arrangement involving cognate programmes at another higher education provider

viii. the succession of an external examiner by a colleague from the examiner's home department and provider

ix. the appointment of more than one external examiner from the same department of the same higher education provider.

Terms of office

c) The duration of an external examiner's appointment will normally be for four years, with an exceptional extension of one year to ensure continuity.

d) An external examiner may be reappointed in exceptional circumstances but only after a period of five years or more has elapsed since their last appointment.

e) External examiners normally hold no more than two external examiner appointments for taught programmes/modules at any point in time.

6. Degree-awarding bodies ensure that all external examiners they appoint are informed about organisational procedures, practices, and academic regulations, and the crucial value of external examiners' feedback to the degree-awarding body as part of the broader system of quality assurance and enhancement.
7. Degree-awarding bodies communicate clearly in writing to all concerned the: modules, programmes and/or qualifications(s) to which each external examiner is appointed various roles, powers and responsibilities assigned to their external examiners, including the extent of their authority in examination boards.
8. Higher education providers include the name, position and institution of their external examiners in module or programme information provided to students.
9. Prior to the confirmation of mark lists, pass lists or similar documents, degree-awarding bodies expect external examiners to endorse the outcomes of the assessment processes they have been appointed to scrutinise.
10. Degree-awarding bodies provide external examiners with sufficient evidence to enable them to discharge their responsibilities.
11. Higher education providers recognise the importance, and mutual benefit, of the work undertaken by many of their staff as external examiners for other providers and agree with staff the time they need to fulfil these duties.
12. External examiners submit a report annually, at a time determined by the degree-awarding body, to the head of the degree-awarding body or to one or more named individuals that he/she designates.
13. External examiners' annual reports provide clear and informative feedback to the degree awarding body on those areas defined for the role in Indicators 2 and 3 (the core content). In addition, their reports: confirm that sufficient evidence was received to enable the role to be fulfilled (where evidence was insufficient, they give details) state whether issues raised in the previous report(s) have been, or are being, addressed to their satisfaction

address any issues as specifically required by any relevant professional body give an overview of their term of office (when concluded).

14. Higher education providers make external examiners' annual reports available in full to students, with the sole exception of any confidential report made directly, and separately, to the head of the degree-awarding body.
15. At both organisational and subject/programme level, degree-awarding bodies give full and serious consideration to the comments and recommendations contained in external examiners' reports. The actions taken as a result of reports, or the reasons for not taking action, are formally recorded and circulated to those concerned. Degree-awarding bodies ensure that student representatives are given the opportunity to be fully involved in this process, enabling them to understand all the issues raised and the degree-awarding body's response. At organisational level the general issues and themes arising from the reports are reviewed.
16. Degree-awarding bodies provide external examiners with a considered and timely response to their comments and recommendations, outlining any actions they will be taking as a result of the reasons for not taking action.
17. Degree-awarding bodies inform external examiners, in writing at the beginning of their term of office, that they have a right to raise any matter of serious concern with the head of the degree-awarding body, if necessary by means of a separate confidential written report. Degree-awarding bodies provide a considered and timely response to any confidential report received, outlining any actions they will be taking as a result.
18. Where an external examiner has a serious concern relating to systemic failings with the academic standards of a programme or programmes and has exhausted all published applicable internal procedures, including the submission of a confidential report to the head of the degree-awarding body, he/she may invoke *QAA's concerns scheme* or inform the relevant professional, statutory or regulatory body.

Expectation B8 (Programme monitoring and review)

Higher education providers, in discharging their responsibilities for setting and maintaining academic standards and assuring and enhancing the quality of learning opportunities, operate effective, regular and systematic processes for monitoring and for review of programmes.

Indicators

1. Higher education providers maintain strategic oversight of the processes for, and outcomes of, programme monitoring and programme review, to ensure processes are applied systematically and operated consistently.
2. Higher education providers take deliberate steps to use the outcomes of programme monitoring and review processes for enhancement purposes.
3. Higher education providers operate a process to protect the academic interests of students when a programme is closed.
4. Higher education providers define processes, roles and responsibilities for programme monitoring and programme review and communicate them to those involved.
5. Higher education providers evaluate their processes for programme monitoring and review and take action to improve them where necessary.
6. Higher education providers make use of reference points and draw on expertise from those outside the programme in their processes for programme monitoring and review.
7. Higher education providers involve students in programme monitoring and review processes.
8. Higher education providers enable staff and other participants to contribute effectively to programme monitoring and programme review by putting in place appropriate arrangements for their support and development.

Expectation B9 (Academic appeals and student complaints)

Higher education providers have procedures for handling academic appeals and student complaints about the quality of learning opportunities; these procedures are fair, accessible and timely, and enable enhancement.

Indicators

1. Higher education providers make available opportunities for students to raise matters of concern without risk of disadvantage.
2. Higher education providers have procedures which encourage constructive engagement with the appeals and complaints process and which offer opportunities for early and/or informal resolution.
3. Higher education providers have accessible appeals and complaints procedures.
4. Clear and accurate advice and guidance is available for students making an appeal or complaint, and for staff involved in handling or supporting appeals and complaints.
5. Academic appeals and complaints procedures are conducted in a timely and fair manner.

6. Higher education providers ensure that appropriate action is taken following an appeal or complaint.
7. Higher education providers monitor and evaluate the effectiveness of their appeals and complaints procedures, and reflect on the outcomes of those procedures for enhancement purposes.

Expectation B10 (Managing higher education provision with others)

Degree-awarding bodies take ultimate responsibility for academic standards and the quality of learning opportunities, irrespective of where these are delivered or who provides them. Arrangements for delivering learning opportunities with organisations other than the degree awarding body are implemented securely and managed effectively.

Indicators

1. A strategic approach to delivering learning opportunities with others is adopted. Appropriate levels of resources (including staff) are committed to the activities to ensure that the necessary oversight is sustained.
2. Governance arrangements at appropriate levels are in place for all learning opportunities which are not directly provided by the degree-awarding body. Arrangements for learning to be delivered, or support to be provided, are developed, agreed and managed in accordance with the formally stated policies and procedures of the degree-awarding body.
3. Policies and procedures ensure that there are adequate safeguards against financial impropriety or conflicts of interest that might compromise academic standards or the quality of learning opportunities. Consideration of the business case is conducted separately from approval of the academic proposal.
4. Degree-awarding bodies that engage with other authorised awarding bodies to provide a programme of study leading to a joint academic award satisfy themselves that they have the legal capacity to do so.
5. The risks of each arrangement to deliver learning opportunities with others are assessed at the outset and reviewed subsequently on a periodic basis. Appropriate and proportionate safeguards to manage the risks of the various arrangements are determined and put in place.
6. Appropriate and proportionate due diligence procedures are determined for each proposed arrangement for delivering learning opportunities with an organisation other than the degree-awarding body. They are conducted periodically to check the capacity of the other organisation to continue to fulfil its designated role in the arrangement.

7. There is a written and legally binding agreement, or other document, setting out the rights and obligations of the parties, which is regularly monitored and reviewed. It is signed by the authorised representatives of the degree-awarding body (or higher education provider without degree-awarding powers arranging provision by a third party) and by the delivery organisation, support provider or partner(s) before the relevant activity commences.
8. Degree-awarding bodies take responsibility for ensuring that they retain proper control of the academic standards of awards where learning opportunities are delivered with others. No serial arrangements are undertaken without the express written permission of the degree-awarding body which retains oversight of what is being done in its name.
9. Degree-awarding bodies retain responsibility for ensuring that students admitted to a programme who wish to complete it under their awarding authority can do so in the event that a delivery organisation or support provider or partner withdraws from an arrangement or that the degree-awarding body decides to terminate an arrangement.
10. All higher education providers maintain records (by type and category) of all arrangements for delivering learning opportunities with others that are subject to a formal agreement.
11. Degree-awarding bodies are responsible for the academic standards of all credit and qualifications granted in their name. This responsibility is never delegated. Therefore, degree-awarding bodies ensure that the standards of any of their awards involving learning opportunities delivered by others are equivalent to the standards set for other awards that they confer at the same level. They are also consistent with UK national requirements.
12. When making arrangements to deliver a programme with others, degree-awarding bodies fulfil the requirements of any professional, statutory and regulatory body (PSRB) that has approved or recognised the programme or qualification, in relation to aspects of its delivery and any associated formal agreements. The status of the programme or qualification in respect of PSRB recognition is made clear to prospective students.
13. Degree-awarding bodies approve module(s) and programmes delivered through an arrangement with another delivery organisation, support provider or partner through processes that are at least as rigorous, secure and open to scrutiny as those for assuring quality and academic standards for programmes directly provided by the degree-awarding body.

14. Degree-awarding bodies clarify which organisation is responsible for admitting and registering a student to modules or programmes delivered with others, and ensure that admissions are consistent with their own admissions policies.
15. Degree-awarding bodies ensure that delivery organisations involved in the assessment of students understand and follow the assessment requirements approved by the degree-awarding body for the components or programmes being assessed in order to maintain its academic standards. In the case of joint, dual/double and multiple awards or for study abroad and student exchanges, degree-awarding bodies agree with their partners on the division of assessment responsibilities and the assessment regulations and requirements which apply.
16. Degree-awarding bodies retain ultimate responsibility for the appointment, briefing and functions of external examiners. The external examining procedures for qualifications where learning opportunities are delivered with others are consistent with the degree awarding body's approved practices.
17. Degree-awarding bodies ensure that modules and programmes offered through other delivery organisations, support providers or partners are monitored and reviewed through procedures that are consistent with, or comparable to, those used for modules or programmes provided directly by them.
18. Degree-awarding bodies ensure that they have effective control over the accuracy of all public information, publicity and promotional activity relating to learning opportunities delivered with others which lead to their awards. Information is produced for prospective and current students which is fit for purpose, accessible and trustworthy. Delivery organisations or support providers are provided with all information necessary for the effective delivery of the learning or support.
19. When degree-awarding bodies make arrangements for the delivery of learning opportunities with others, they ensure that they retain authority for awarding certificates and issuing detailed records of study in relation to student achievement. The certificate and/or record of academic achievement states the principal language of instruction and/or assessment where this is not English.22 Subject to any overriding statutory or other legal provision in any relevant jurisdiction, the certificate and/or the record of achievement records the name and location of any other higher education provider involved in the delivery of the programme of study.23 Where information relating to the language of study or to the name and location of the delivery organisation or partner is

recorded on the record of achievement only, the certificate refers to the existence of this formal record.

Expectation B11 (Research degrees)

Research degrees are awarded in a research environment that provides secure academic standards for doing research and learning about research approaches, methods, procedures and protocols. This environment offers students quality of opportunities and the support they need to achieve successful academic, personal and professional outcomes from their research degrees.

Indicators

1. Higher education providers that are research degree awarding bodies have regulations for research degrees that are clear and readily available to research students and staff, including examiners. Where appropriate, regulations are supplemented by similarly accessible, subject-specific guidance at the level of the faculty, school, department, research centre, or research institute.
2. Higher education providers develop, implement and keep under review codes of practice for research degrees, which are widely applicable and help enable the higher education provider meet the Expectation of this Chapter. The codes are readily available to all students and staff involved in research degrees, and written in clear language understood by all users.
3. Higher education providers monitor their research degree provision against internal and external indicators and targets that reflect the context in which research degrees are being offered.
4. Higher education providers accept research students only into an environment that provides support for doing and learning about research, and where excellent research, recognised by the relevant subject community, is occurring.
5. Higher education providers' admissions procedures for research degrees are clear, consistently applied and demonstrate equality of opportunity.
6. Only appropriately qualified and prepared applicants are admitted to research degree programmes. Admissions decisions involve at least two members of the higher education provider's staff who have received training and guidance for the selection and admission of research degree students. The decision-making process enables the higher education provider to assure itself that balanced and independent admissions decisions have been made in accordance with its admissions policy.
7. Higher education providers define and communicate clearly the responsibilities and entitlements of students undertaking research degree programmes.

8. Research students are provided with sufficient information to enable them to begin their studies with an understanding of the environment in which they will be working.
9. Higher education providers appoint supervisors with the appropriate skills and subject knowledge to support and encourage research students, and to monitor their progress effectively.
10. Each research student has a supervisory team containing a main supervisor who is the clearly identified point of contact.
11. Higher education providers ensure that the responsibilities of research student supervisors are readily available and clearly communicated to supervisors and students.
12. Higher education providers ensure that individual supervisors have sufficient time to carry out their responsibilities effectively.
13. Higher education providers put in place clearly defined mechanisms for monitoring and supporting research student progress, including formal and explicit reviews of progress at different stages. Research students, supervisors and other relevant staff are made aware of progress monitoring mechanisms, including the importance of keeping appropriate records of the outcomes of meetings and related activities.
14. Research students have appropriate opportunities for developing research, personal and professional skills. Each research student's development needs are identified and agreed jointly by the student and appropriate staff at the start of the degree; these are regularly reviewed and updated as appropriate.
15. Higher education providers put in place mechanisms to collect, review and respond as appropriate to evaluations from those concerned with research degrees, including individual research students and groups of research students or their representatives. Evaluations are considered openly and constructively and the results are communicated appropriately.
16. Higher education providers that are research degree awarding bodies use criteria for assessing research degrees that enable them to define their academic standards and the achievements of their graduates. The criteria used to assess research degrees are clear and readily available to research students, staff and examiners.
17. Research degree final assessment procedures are clear and are operated rigorously, fairly, and consistently. They include input from an external examiner and are carried out to a reasonable timescale. Assessment procedures are communicated clearly to research students, supervisors and examiners.

18. Higher education providers put in place and promote independent and formal procedures for dealing with complaints and appeals that are fair, clear to all concerned, robust, and applied consistently. The acceptable grounds for complaints and appeals are clearly defined.

Part C: Information about Higher Education Provision

Expectation C

Higher education providers produce information for their intended audiences about the learning opportunities they offer that is fit for purpose, accessible and trustworthy.

Indicators

1. Higher education providers publish information that describes their mission, values and overall strategy.
2. Higher education providers describe the process for application and admission to the programme of study.
3. Higher education providers publish information to help prospective students select their programme with an understanding of the academic environment in which they will be studying and the provision that will be made to enable their development and achievement.
4. Information on the programme of study is issued to current students at the start of their programme and throughout their studies.
5. Higher education providers set out what they expect of current students and what current students can expect of the higher education provider.
6. When students leave their programme of study, higher education providers issue to them a detailed record of their studies, which gives evidence to others of the students' achievement in their academic programme.
7. Higher education providers:
    - set out their arrangements for managing academic standards and quality assurance and enhancement and describe the data and information used to support its implementation
    - maintain records (by type and category) of all arrangements for delivering higher education with others that are subject to a formal agreement.

3. Risk-Based Approaches

The purpose of this chapter is to explore 'risk-based' approaches to regulation. It defines 'risk-based regulation' and examines its evolution, merits, and limitations. This is followed by an examination of the different ways in which a risk-based approach can be implemented and a review of the literature concerning risk-based approaches to quality assurance in higher education.

3.1. Defining Risk-Based Regulation

Risk-based regulation is built upon the allocation of regulatory resource in proportion to the risks, calculated as the product of the perceived impact and likelihood of their occurrence, posed to the regulator's objectives (Black, 2005; Rothstein *et al.*, 2006b). Risk-based regulation ostensibly provides practitioners with a means with which to "maximise the benefits of regulation while minimising the burdens on regulatees by offering 'targeted' and 'proportionate' interventions" (Rothstein *et al.*, 2006a, p.97). For QAA, this means the ability to lighten or eliminate reviews for low-risk providers, leaving them free to prosper, whilst using the resource saved to conduct reviews of high-risk providers and quickly eliminate or prevent any poor practice, all at the same or reduced cost to the taxpayer. In theory, everybody wins.

Regulators have always had to prioritise their limited resource however; the issue of overburdened regulators is not a new one (Pontell, 1978). QAA for example is responsible for assuring the quality and standards of over 700 higher education providers in the UK and overseas (HEFCE, 2015c), direct regulation of the 'Access to Higher Education' qualification, advising the privy council on applications for degree awarding powers and the use of the university title, and contributes towards decisions resulting in providers being awarded 'highly trusted status' by the UK Borders Agency. It must fulfil its responsibilities across four different devolved legislatures (more if one considers UK higher education provided overseas) and within a limited budget of £16M (QAA, 2015b). Risk-based approaches are intended to move this implicit prioritisation of resource forward with the explicit determination of risk through assessment frameworks and help frame solutions not just in terms of costs and benefits, but also the impact and probability of future uncertainties in a "politically compelling" manner (Black, 2005, p.4).

Risk-based approaches have evolved from their original focus on inspection planning (Hampton, 2005) to cover the full purview of regulatory duties. Risk-based approaches to enforcement, standard setting and licencing continue to be developed, such as the CQC's risk-based approach

to registering health and social care providers legally permitting them to practice controlled procedures (CQC, 2015a), as the growth in regulators' responsibilities is not matched by available resource. In line with *Students at the Heart of the System* (BIS, 2011) and the ongoing debate in higher education, this thesis focuses on the risk-based prioritisation of inspections.

3.2. The Emergence of Risk-Based Regulation

Regulation has become pervasive in modern society. Rising international competition, increasing economic integration and the embracing of (quasi) markets has resulted in governments transforming from providers of macroeconomic stability and 'merit goods' to guardians against market failures (Loughlin and Scott, 1997; Majone, 1994). Majone argues that the increased number of EU directives and necessary harmonisation of its members' laws, the privatisation of state monopolies now requiring price and competition controls, and the introduction of internal markets with public providers and commissioners operating through contractual arrangements have all led to a proliferation of regulation (Majone, 1997). This shift from the 'interventionist' to the 'standard-setting and enforcement' government has heralded a reduction in central administration and a move towards the establishment of specialised arms-length regulatory bodies. These bodies are not only tasked with the oversight of private actors' adherence to competition rules in newly created markets, but also the financial management and quality of providers in the traditionally 'public' spheres such as healthcare and education (Hood *et al.*, 1999).

With the rise of what has become known as the 'regulatory state', and the corresponding rise of New Public Management (Hood, 1995), Black (2005) suggests we are now experiencing 'New Public Risk Management'. As the virtues of private business practices were extolled during the 1980s and 1990s, a strong deregulatory rhetoric developed. In Europe and America regulators were accused of stifling growth and burdening industry with 'red tape' (Majone, 1990; Breyer *et al.*, 1999; Hutter, 2005) and, following a spate of privatisations, the Major Government in the UK passed the *Deregulation and Contracting out Act* (1994).

In response to concerns over burden and earlier regulatory failings, the mid-1990's saw the popularisation of *enforced self-regulation* (Ayres and Braithwaite, 1992; Parker, 2002). Adopted in diverse fields including health and safety and the quality assurance of higher education, regulators ceded responsibility for assessing compliance with standards to regulatees seen as better placed to manage technical issues and instead audited the management systems of regulatees' developed to ensure their own compliance. Self-regulatory approaches were seen as the answer

to the inherent limitations of hierarchical oversight, and were cheaper too (Lodge and Wegrich, 2012).

The evolution of regulatory practices was continued by the Blair Government with the establishment of the *Better Regulation Task Force* (BRTF) and a plethora of independent regulatory agencies (see for example Thatcher, 2002; Patterson and Lilburne, 2003). The Better Regulation Task Force published its 'Principles of Good Regulation' in 1997 (BRTF, 1997) advocating risk-based regulatory approaches which built upon enforced self-regulation (Lodge and Wegrich, 2012). In the following years the National Audit Office (NAO) and the Treasury followed suit between them recommending risk-based approaches for, amongst others, the Gaming Board, Maritime and Coast Agency, Occupational Pensions Regulatory Authority, Ofsted, the Social Care sector and efficiency across government (NAO, 2000, 2001, 2002, 2003; BRTF, 2002, 2004; Gershon, 2004).

It was the publication of *Reducing administrative burdens: effective inspection and enforcement* (Hampton, 2005) - the 'Hampton Report' - which cemented the use of risk-based approaches amongst oversight bodies. Drawing on evidence from the long-standing practice in food safety and occupational health and safety regulation, the report emphatically endorsed the use of risk-based approaches for all UK regulators and gave rise to the ubiquitous 'Hampton principles'. Later that year the then Chancellor Gordon Brown continued to laud risk-based approaches telling the Confederation of British Industry (CBI):

> "In the old regulatory model – and for more than one hundred years – the implicit principle from health and safety to the administration of tax and financial services has been, irrespective of known risks or past results, 100 per cent inspection whether it be premises, procedures or practices.
>
> This approach, followed for more than a century of regulation by governments of all parties is outdated. The better, and in my opinion the correct, modern model of regulation – the risk based approach – is based on trust in the responsible company, the engaged employee and the educated consumer, leading government to focus its attention where it should: no inspection without justification, no form filling without justification, and no information requirements without justification, not just a light touch but a limited touch"
>
> (Brown, 2005).

The following year the *Better Regulation Commission* (BRC) was established "to advise the Government on action to reduce unnecessary regulatory and administrative burdens, and ensure

that regulation and its enforcement are proportionate, accountable, consistent, transparent and targeted" (BRC, 2006). The institutionalisation of risk-based approaches was completed one year later when, following the publication of the *Statutory Code of Practice for Regulators*, it became a legal requirement in the UK for all regulators to develop and operate risk-based frameworks (BERR, 2007). As Rothstein *et al.* (2011) asserted, the regulation *of* risk became regulation *by* risk.

High-profile failings, epitomised by the Financial Services Authority (FSA) and the financial crisis of 2008 (see for example Lodge and Wegrich, 2012), have led to a greater scrutiny of risk-based regulation, but have not slowed its proliferation. Recent years have seen risk-based approaches lauded by The Organisation for Economic Co-operation and Development (OECD, 2010) and introduced or revised by *inter alia* the Care Quality Commission (CQC, 2013a, 2015a), Australia's Tertiary Education Quality Standards Agency (TEQSA, 2012, 2015), NHS Improvement (NHS Improvement, 2015), and of course QAA (QAA, 2013a, 2013b). The perceived merits of a risk-based approach responsible for its proliferation are examined below.

3.3. The Merits of a Risk-Based Approach

Risk-based approaches would not have become so popular had they not satisfied, or at least been perceived to have satisfied, the key regulatory requirements of their time. *First* and foremost, risk-based approaches promised a more efficient use of resource. To use the QAA as an example, until the recent introduction of the *HER* approach all universities were reviewed every six years regardless of their prior performance or the available intelligence. Investing resource in reviewing a university that always demonstrated the highest standards and current data suggests is continuing to do so whilst waiting six years to re-review a poorly-performing university that is showing no obvious signs of improvement is arguably a poor use of resource. By prioritising these universities according to the risk posed to the QAA's objectives, the high-performing universities will be rewarded for their efforts by a reduced burden from review whilst the poorly-performing universities will receive the attention they warrant, and the support they need, to improve, all with the same amount of overall resource as before.

Second, the ostensibly objective risk assessments underpinning the risk-based approaches promised regulators a defensible rationale for their prioritisation activity (Sunstein, 2002). Regulators frequently have more to do, and issues to respond to, than resources permit (Black, 2005). Regulators , therefore, cannot and do not prevent all harms. This can be easily exemplified: major healthcare failings at Bristol Royal Infirmary and Mid-Staffordshire hospital saw the unnecessary deaths of hundreds of patients (Healthcare Commission, 2009); confidence in food

safety has been harmed by the BSE, Foot and Mouth disease and e-coli crises; and train crashes at Potters Bar, Ladbroke Grove and Hatfield killed 42 and injured 634 people in a 30-month period (Cullen, 2002; ORR, 2006; Rail Safety and Standards Board, 2005). The seemingly continuous stream of failures and a decline in confidence over what regulation can achieve at the same time as the adoption of New Public Management practices has seen increased oversight and accountability for regulators (Löfstedt, 2008). Conceiving of harms, both to those they are charged with protecting and to themselves, in terms of risks to be managed allows regulators to demonstrate a rational defence for their actions (Rothstein *et al.*, 2006a). Where low-level 'probabilistic' risks have a high public salience, or where the regulator failed to prevent an incident of which the available data contained no forewarning, a demonstrable defence for inaction in the form of a risk assessment is highly desirable and has the further benefit of providing transparency in an era of heightened public accountability (Rothstein *et al.*, 2006a; Lodge and Wegrich, 2012; Demeritt *et al.*, 2015).

Third, the adoption of *en vogue* methodologies provides regulators with legitimacy and aids bureaucratic survival (Meyer and Rowan, 1977). Risk-based approaches promised regulators not only the ability to rationalise, manage and control the notoriously complex challenges of regulation, but the ability to do so whilst appearing forward-thinking in a period of political enamourment with private-sector practices (Hutter, 2005; Baldwin *et al.*, 2012).

The appeal of risk-based approaches is clear: practitioners make rational and efficient use of their resource, they do so in a readily-defensible manner, appear favourable to Government and allow compliant actors to prosper whilst affording non-compliant actors the attention they require. As always, however, the devil is in the detail of the regulator's risk assessment tools and the technical, legal and political implementation of their approach.

3.4. The Limitations of a Risk-Based Approach

There are significant challenges facing the successful implementation of a risk-based approach to regulation (Rothstein *et al.*, 2006b). *First*, the epistemic challenge of correctly determining the actuarial risk of regulatory harms occurring. *Second*, the normative challenge of managing the acceptability of risk to different groups *ex ante* and *ex post*. *Third*, the institutional challenges of operating a risk-based approach once risks have been determined. Each challenge is explored in turn below.

3.4.1. Epistemic Challenges

A risk-based approach to regulation cannot work without an effective method for determining risk. Before any data is entered into a risk model subjective decisions must be made about, amongst other things, what risks are to be assessed, what data should be used, how it will be weighted and aggregated, how 'impact' will be determined, and how much confidence can be placed in the data? Even with these questions answered and with the risk model in place regulators must still decide how much discretionary judgement to allow when considering the output to balance consistency with flexibility. In short, whilst a regulator's risk model may provide an 'objective' output indicating the level of risk, a great deal of subjective decision-making is required to develop the model, interpret the output and determine the correct response (Slovic, 1992).

The collection and use of data also poses a number of challenges. Risk models are often developed to utilise the data that is available rather than the information which would be of most use (OECD, 2010). This is in part driven by the fact that the imposition of additional data collections adds 'regulatory burden' on regulatees, the antithesis of risk-based regulation (see for example Dow and Braithwaite, 2013). Risk assessments necessarily based on limited data due to availability constraints are unlikely to reach the correct conclusion (Kahneman, 2011; Miller, 1962). Where regulatees are aware that the information they provide may be used to prioritise or sanction them, that information will cease to be useful for regulatory purposes (Goodhart, 1984) and cooperation between the regulator and regulatees will be negatively impacted (Baldwin *et al.*, 2012).

Even where extensive, high-quality data are available, risk models face challenges. For example, Turner (1994) highlights the disputed technical underpinnings of quantitative risk assessments, Cohen (1996) and Toft (1996) challenge the higher-level efficacy of the reduction and aggregation of data, whilst Taleb (2010) demonstrates the risk model's inability to account for extreme outlier events, or 'black swans', such as the September 11th 2001 terrorist attacks. More broadly, Reason (1990) and King (2014a) have questioned the ability of models to predict human behaviour. Whatever the prevailing reason, the sole study to date empirically assessing a regulator's risk prioritisation tool has found it actively misleading. The CQC would have been better off doing the opposite of what their 'Intelligent Monitoring' system for prioritising hospital inspections suggested (Griffiths *et al.*, 2016). Finally, risk models focusing on prioritising the inspection and enforcement of individual regulatees may be blind to emerging systematic risks, either because they fall outside of the existing risk model's parameters or because they are affecting all regulatees, including those deemed 'low risk' and spared inspection. Such dangers were foreseen

by Hampton (2005) who suggested the ongoing use of random inspection to continually evaluate regulators' approaches and spot emerging threats across high and low risk providers alike. This however failed to prevent the financial crisis of 2008.

3.4.2. Normative Challenges

Setting aside the challenges to successfully assessing risk, the rational defence for regulatory decision-making offered by an objective risk assessment is likely to be challenged by political considerations. The perceived 'intolerable' risks which concern the general population often differ from 'true', 'probabilistic' risks which a rational regulator should prioritise (Slovic, 1992; Starr, 1969). Government ministers however, elected to represent the population and keen to maintain their careers, may seek to direct regulatory activity towards low-risk 'intolerable' concerns (Breyer, 2009). This was exemplified in 2012 by then Health Secretary's instruction that the CQC undertake inspections of all abortion clinics in England following media revelations about doctors pre-signing second-opinion forms. Whilst a valid concern, these inspections cost £1,000,000 to conduct at short notice and resulted in the cancellation of 580 pre-planned inspections of hospitals and care homes that posed a far greater risk (BBC, 2012). Conversely, whilst espousing the virtues of risk-based approaches, politicians may be unwilling to accept failures of any size that are inevitable when managing risks rather than trying to prevent all harms. Explaining to the family of a vulnerable care home resident who died as a result of neglect that their case was, when probability and consequence are considered, low impact and therefore of limited interest to the regulator is not straightforward. Impact can be very subjective (Beaussier *et al.*, 2016).

'Risk colonisation' means that regulators themselves can also be guilty of prioritising lesser risks with a higher salience in an exercise in bureaucratic self-preservation. Rothstein *et al.* (2006a) cite the example of train safety where regulators feel under pressured to allocate more resource towards the prevention of low-probability, high-casualty accidents which attract media attention and intense lobbying than to high-probability, individual-fatality incidents which total more deaths overall. As discussions concerning nuclear power have shown, attempts to resolve the friction between 'intolerable' and the more 'probabilistic' risks via public education are unlikely to be successful (Douglas, 1992). Some regulators, such as the UK Pensions Regulator, have attempted to accommodate this by including 'loss of public confidence' as a criterion of their risk assessment. The clear danger however is that this undermines the 'rational' tenet of risk-based regulation (Baldwin *et al.*, 2012). Even when regulators are able to successfully deal with the disjuncture between 'intolerable' and 'probabilistic' risks, the public's perception of risk can change much quicker than established regulatory frameworks.

Individual inspectors also face their own normative challenges. If faced with two similar organisations and the performance of the first has led to it being an 'amber' risk whilst nothing is known about the performance of the second: which should be prioritised? Likewise, where no overall aggregated risk score exists, but rather a number of lower-level risk-category scores, should an organisation with two 'amber' risks be prioritised over another with one 'red' risk? It is unlikely simple aggregation rules for broad risk assessment categories will be able to fully deal with the nuances of such challenges.

3.4.3. Institutional Challenges

Successfully allocating resource in proportion to the risks to one's objectives may not only be constrained by epistemic challenges, but also by a regulator's internal, institutional challenges (Lodge and Wegrich, 2012). Regulators can have wide ranging and often contradictory objectives. The NHS finance and governance regulator *Monitor* for example is responsible both for promoting competition and protecting the financial health of NHS Foundation Trusts. High barriers of entry to the market for new private providers pose a risk to increased competition yet reduce the risk of NHS Foundation Trusts running into financial difficulty (PAC, 2014).

Even when the risk regulatees pose is clear, regulators allocating their resources in accordance to those risks cannot be sure it will provide the best use of resource. In practice the resource required to rectify non-compliance is not uniform across regulatees but might vary, *inter alia* in accordance with their cultures, attitudes and capacities. The amount of resource required to mitigate the risk from a recalcitrant regulatee may far exceed the resource required to educate numerous regulatees who wish to comply and simply require the necessary information or education to allow them to do so. Any attempt to account for this by the regulator may however prove counterproductive as it rewards less amenable regulatees with a lower risk score than they merit.

Black and Baldwin (2010) further highlight that a standard risk-based approach will not instruct practitioners in the best approach to take towards individual regulatees who may respond better to different enforcement approaches. For example, risk-based regulators are unlikely to prioritise the education and persuasion approach which mitigates the risks posed by a number of smaller bodies as their size will limit the impact they can have and therefore the risk they pose. Where the non-compliance of smaller, lower impact, bodies is not prioritised, however, issues can progressively escalate across a large number of regulatees and develop into a systematic risk.

Regulators may also face difficulties acting on an identified risk with any great speed as they are only one actor in part of a wider regulatory regime (FSA, 2009). Regulatory action often requires a co-ordinated, multiagency response. In English healthcare, for example, persistent failings

including high death rates at a Foundation Trust may require action from the CQC as the quality regulator, Monitor as the governance regulator and body with the ultimate power to impose change, and the General Medical Council and/or the Royal College of Surgeons to manage the individual medics.

On an individual level, regulatory staff can also struggle with the implementation of a risk-based approach. Not only must regulators overcome the challenges of obtaining sufficient data, including the capture of staff members' tacit knowledge, and the effective calculation of complex regulatory risks, but they must then present this in a way simple enough for front-line staff to successfully interpret. Staff can find it a difficult transition in culture from the absolute prevention of failure and 'ticking boxes' to the management of risks of failure, especially the shift away from all failures being unacceptable and someone's fault (Douglas, 1992; OECD, 2010). Black (2005) highlights that staff in such circumstances, especially those not enamoured with the new risk-based approach, may be tempted to reverse engineer risk assessments to ensure a result consistent with their perception rather than the probabilistic score from a model. When the risks identified by a regulator are clustered, either geographically or by area of expertise, the ability to prioritise all these risks may be constrained by the mobility and fungibility of the workforce. Finally, risk-based approaches require a substantial analytical underpinning in order to source, manage, interpret, score, aggregate and disseminate information on the risks posed by providers. With finite resource, this capability must come at the expense of another aspect of the regulator's functions (Lodge and Wegrich, 2012).

In summary, risk-based approaches theoretically offer regulators an efficient means of operation in austere, deregulatory times and a rational defence when challenged over their decision making. In practice their implementation is constrained by *inter alia* political interference, immovable public perceptions, ill-defined or contradictory regulatory frameworks, the differing attitudes of regulatees, and resource limitations. These constraints are secondary however to the epistemic challenges. If a regulator cannot assess risk, a risk-based approach will be fundamentally flawed and the subsequent operational challenges are somewhat moot. It is this regulatory challenge of assessing risk which this thesis focuses on.

3.5. Variations of Risk-Based Approaches

Few 21st century regulatory strategy documents have escaped the platitudes of risk-based approaches (see for example Australian Skills Quality Authority, 2016; Bar Standards Board, 2016;

CQC, 2015a; Ofsted, 2015a; SRA, 2014; TEQSA, 2012, 2015). Regulators, however, rarely detail *how* their approach is risk based. The Care Quality Commission's experience with its high-profile 'Intelligent Monitoring' approach shows that there are good reasons for their reticence. In 2014, following the publication of over 8,000 'Intelligent Monitoring' reports risk rating each GP in the country based on available metrics, CQC faced an understandable backlash from GPs many of whom were indignant at being publicly stigmatised by a regulator that had not even set foot in their premises (Lind, 2014). Relations with GPs were severely damaged and a vote of no confidence in the Chief Inspector of Primary Care was passed by the Royal College of General Practitioners (RCGP, 2015). There was genuine concern that patients could be deterred from seeking necessary treatment (Millett and Bostock, 2014). To compound matters, the publication of the individual 'Intelligent Monitoring' risk reports allowed the BBC to identify that the controversial risk ratings had been calculated incorrectly (Bloch, 2014). Finally, the publication of the 'Intelligent Monitoring' risk reports for NHS hospitals allowed the risk ratings to be compared with the subsequent inspection findings revealing that the risk ratings were worse at prioritising inspections than random selection (Griffiths *et al.*, 2016). Despite regulators' generally limited disclosure of the specifics of their risk-based approaches, the approaches can generally be placed into one of three broad categories. These three broad categories are explored below.

3.5.1. Rules-Based Approaches

Typically utilised by regulators developing their first risk-assessment approach, regulatees can be assigned one of a small number of risk categories by means of a simple, and often contextual, rules-based assessment. This is exemplified by Maritime and Coastguard Agency's simple, rules-based approach based on ship type, age and inspection history (NAO, 2009). Similarly, QAA's *Higher Education Review* approach differentiates between higher and lower risk providers on the basis of compliance history, complaints received, and whether they have "undergone significant material change" (QAA, 2013a, 53). Inspectors from the Food Standards Agency score businesses on the hazards present and how willing and able they are to manage them, and then use that score to calculate the suitable length of time until the next inspection (Food Standards Agency, 2016). The Environment Agency's system requires regulatees to complete risk assessments themselves based on a series of categorical questions. The result is a transparent calculation which determines both the regulatee's licence fee and their risk-rating which in turn determines their inspection frequency (Environment Agency, 2014).

Such contextual, rules-based approaches have clear advantages. The simplicity of the approach means little data or analysis is required saving on staffing and infrastructure costs. The rules-based

approach also eliminates many of the epistemic challenges of a risk-based approach by simply using little if any data. The use of contextual data over performance data also makes such systems hard to game: one cannot readily change a fishing trawler into a canal boat, or amend the result of a historical QAA review.

The simplicity of such approaches does create a number of issues however. *First*, they are not very discerning and, arguably, unfairly discriminatory. In 2011, the European Court of Justice ruled it illegal for motor insurers to charge drivers more based on their sex (European Court of Justice, 2011). Whilst males cost insurers more in pay-outs than females, being a male doesn't inherently make you a more dangerous driver and, as a characteristic outside of an individual's control, it is unfair to discriminate against them. One could extend this argument to all manner of regulatees. Being a district general hospital may put you in a class of hospital more likely to be non-compliant than a large, urban hospital but the property of being a district general hospital cannot be changed nor does it make an individual district general hospital inherently risky. *Second*, the rules are more likely to be based on flawed assumptions than data-led models. The simple criteria are based on an *a priori* selection of measures rather than a statistical analysis of what best predicts regulatory findings. As will be shown in chapter six, counterintuitively, those providers that have previously been judged 'unsatisfactory' by the QAA are actually more likely to be 'satisfactory' on their subsequent review than those providers that were previously 'satisfactory'. *Third*, rules-based approaches do not fully resolve issues of prioritisation. Once regulatees have been divided into a small number of risk categories, how then are they prioritised within their categories? *Finally*, regulators leave themselves open to challenge when a regulatee deemed 'low-risk' is shown to be non-compliant and it transpires that performance data not included in the simplistic model indicated as much.

Simple, rules-based approaches utilising contextual data are therefore cheap and transparent, but fail to accurately identify individual regulatees of concern and can be unfairly discriminatory. The remaining, and most popular, prioritisation techniques rely heavily on the use of metrics to target individual regulatees. The key difference between these data-reliant processes is how the metrics are selected and the use of 'expert interpretation' in the prioritisation decisions.

3.5.2. Data-Informed Approaches

Data-informed prioritisation tools aggregate a large number of *a priori* metrics to generate a risk rating and or report used to inform, but not replace, expert judgement. A pioneer in the field, Financial Services Authority's Advanced Risk Responsive Operating Framework (ARROW) tool saw 45 'elements', irreducible areas of risk, rated on a four-point scale, either subjectively or

automatically depending on the element, and mapped to seven 'risks to objectives'. The ARROW tool automated the weighted aggregation of the 45 elements and sent the result to a supervisor who had the ability to override any risk rating. The final report was then sent with others to a panel to prioritise regulatory action (FSA, 2003, 2002).

CQC's aforementioned 'Intelligent Monitoring' approach is not dissimilar. Approximately 150 metrics, each selected in consultation with experts and relating to one of CQC's five 'key questions' are each automatically scored on a three-point scale based on deviations from a specified performance target or national average. Each metric is either scored 'no evidence of risk', 'risk', or 'elevated risk' and assigned a score of zero, one or two respectively. An overall risk score is then calculated by summing the assigned metric scores and dividing it by the maximum number of points available for a specific provider. These overall risk scores are calculated simultaneously on a regular basis and are used to rank the providers. The Chief Inspector and their colleagues then prioritise the inspections based on the risk-ranking table and any other evidence that comes to light (CQC, 2014b, 2014a).

Prior to its 'Intelligent Monitoring' approach, CQC operated 'Quality and Risk Profiles' which contained approximately 1,000 metrics for each NHS trust. That is, every metric the CQC could obtain that related to one of 16 'outcomes'. Each metric was automatically scored on a seven-point scale based on each provider's deviation from the national average and then all individual metric scores were aggregated, based on three weighting factors, to form an 'outcome risk estimate' on an eight-point scale for each of the 16 'outcomes' (CQC, 2013c). Rather than a panel, an individual inspector would prioritise their own activity based on the QRPs for each of their 60 or so health and social care providers in their varied portfolio (Griffiths, 2012).

Australia's *Tertiary Education Quality and Standards Agency* (TEQSA) introduced its risk-based approach in 2012 and annually assessed risk by casting 'expert judgement' on a set of 46 quantitative and qualitative metrics selected *a priori* by experts and awarding a 'traffic light' risk score of red, amber or green. Overall judgement was made concerning the 'risk to students', 'risk of provider collapse' and 'risk to sector reputation' and regulatory action prioritised accordingly (TEQSA, 2012). Following strong complaints from the sector about burden and duplication within the system, a review into the regulation of the higher education sector was announced the following year (TEQSA, 2013). TEQSA's need for such an 'elaborate' risk assessment framework and its effectiveness were also questioned (Dow and Braithwaite, 2013). Subsequently, in March 2014, TEQSA published 'a simplified and more robust' regulatory risk framework which comprised an annual review in which each provider is 'holistically' rated red, amber or green using

professional judgement having reviewed the reduced set of 20 metrics, the thresholds for which are not published and are determined subjectively (TEQSA, 2014b, 2014a).

The benefits of prioritising inspections using such 'data-informed' approaches include allowing regulators to: target individual regulatees, present themselves as monitoring all aspects of performance, and provide a rational defence when regulatory failings occur. Risk assessments considering large quantities of data are also difficult for regulatees to game; simply shuffling resource to improve performance on one metric will likely result in worsening performance on another of the large number of metrics. Finally, the attraction of adding a layer of expert interpretation is clear. Taking higher education as an example, HEIs vary tremendously and knowing that, for example, Oxford has fewer contact hours than most due to their system of individual tuition rather than a lack of attention to students, could improve the use of data to target those HEIs with 'unsatisfactory' quality assurance processes. The expert interpretation of the risk scores means regulators are not beholden to the data. Tacit knowledge, notoriously difficult to capture in large-scale quantitative models, can be employed when the data is misleading.

'Data-informed' approaches have three significant failings however. *First*, there are a number of epistemic issues. The numerous and wide-ranging metrics are selected solely by the regulator or in conjunction with interested parties on the *a priori* basis on what parties *believe*, rather than *know empirically*, to predict non-compliance, or they are pressured to include for political purposes. The one peer-reviewed study to date shows CQC's Intelligent Monitoring tool to be actively worse at identifying non-compliance than random selection (Griffiths *et al.*, 2016). The collection of numerous metrics results from investing resource in gathering everything that can be measured, rather than developing new, better metrics, and can lead to any signal arising from one metric being lost in the noise of others. Furthermore, with significant numbers of metrics, even if all performance is normally distributed every regulatee would expect to be flagged as an extreme outlier for at least one metric making extreme performance difficult to interpret. This was the case with the CQC's Quality and Risk Profiles where, with approximately 1,000 metrics, one would expect a perfectly average hospital to be identified as performing 'much worse than expected' (two standard deviations worse than the national average) on 25 metrics by chance alone.

Second, there is a limit to how much information someone can process (Kahneman *et al.*, 1982). Taking CQC's QRP as an example, an inspector with responsibility for 60 health and social care organisations would be faced with 60 sets of 16 risk estimates, 960 in total, each time the risk

assessments were updated. Even ignoring all the additional information contained within each of the 960 risk estimates, this is clearly too much information for an individual to rationally process and make a reasoned, consistent judgement upon. Obtaining additional consistency over 800 inspectors, or multiple prioritisation panels, will add yet more difficulty. A further issue caused by expert interpretation is the constraint it places on the frequency with which prioritisation decisions can be made; physically interpreting thousands of risk assessments is prohibitively expensive and time consuming, reinterpreting risk assessments for each regulatee every time a metric is updated is impossible (Griffiths, 2012; Pollard, 2011).

Third, and of greatest importance, is that, even when there are fewer risk assessments to consider, the well-established empirical literature tells us that the consistency, and accuracy, of expert decisions will be highly problematic. In 1954, Meehl reviewed some 20 studies evaluating the performance of expert decision-making and prediction against simple regression models in diverse fields including higher education, parole violation, pilot training and clinical recidivism and found the models outperformed or equalled experts in every study (Meehl, 1954). The finding that simple models are superior, or at worst equal to, expert judgements has been repeatedly and consistently confirmed in multiple studies in fields as diverse as business bankruptcy (Deakin, 1972; Libby, 1976; Beaver, 1966), survival times (Einhorn, 1972), the outcome of American and English football games (Forrest *et al.*, 2005; Song *et al.*, 2007), police disciplinary matters (Inwald, 1988), military training success (Bloom and Brundage, 1947) heart attacks (Lee *et al.*, 1986; Goldman *et al.*, 1988), neuropsychological and psychiatric diagnosis (Goldberg, 1965; Wedding, 1983; Filskov, 1984), financial auditing (Brown, 1983), future prices of Bordeaux wines (Ashenfelter, 2008), and violence (Werner *et al.*, 1983; Miller and Morris, 1988). Moreover, Grove *et als'* (2000) meta-analysis of 136 studies of clinical judgement versus statistical prediction concluded that "superiority for [statistical]-prediction techniques was consistent, regardless of the judgment task, type of judges, judges' amounts of experience, or the types of data being combined" (Grove *et al.*, 2000, p.19). It has even been demonstrated that models outperform clinicians when the clinicians have the output of the model available to assist their judgement (see for example Goldberg, 1968; Montier, 2009).

Over 200 robust studies into the performance of expert decision-making versus simple statistical models been performed. No convincing exception has been reported (Kahneman, 2011)[2]. Strikingly, Dawes (1979) found that 'improper' models, those that have no weighting (β regression

[2] Meehl (1965) concluded that one individual study did show clinical judgement to be superior; however, this case is widely disputed (Goldberg, 1968, Daves et al. 1988, Kahnemann, 2011).

coefficients), outperform expert judgement in many fields. Decades after his original findings and following their repeated confirmation, Meehl has concluded "There is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one" (Meehl, 1986, p.4).

The reasons for, and extent of, models outperforming expert decision-making will vary by field and by expert. There are some universal factors which constrain decision making however. Tetlock and Gardner's (2016) comprehensive *Good Judgement Project* demonstrated experts were consistently overconfident about the completeness of the information available to them and were subject to multiple illusions and self-serving biases such as group think, the will to agree with superiors, and to avoid absolute judgements (Janis, 1982; Reason, 1990; Mellers *et al.*, 2015). The use of expert judgement to prioritise QAA reviews would be further confounded by the regulatory environment. Decisions to prioritise reviews would be made in a 'low validity' environment in which feedback is limited (it is unlikely experts will discover when they have failed to prioritise a non-compliant regulatee) and 'noise' is substantial (it is not clear whether the non-compliance relates to information on which the prioritisation decision was made) (Kahneman and Klein, 2009; Tetlock and Gardner, 2016). Indeed, it is such environments where simple models perform best against human decision-making; a concern for any risk-based approach built upon expert interpretation (Tetlock and Gardner, 2016; Kahneman, 2011).

'Data informed' approaches therefore tend to comprise a large number of metrics selected *a priori* without statistical assessment and aggregated to provide a risk report and overall risk score for consideration by experts. They allow regulators to consider, and be seen to consider, a wide range of performance measures and incorporate tacit knowledge. These advantages however are outweighed by the significant epistemic challenges, issues with consistency, and the fact that numerous studies have assiduously demonstrated expert interpretation to be worse than simple models.

3.5.3. Data-Driven Approaches

Data-driven approaches make use of machine-learning techniques to identify useful metrics and develop optimal statistical models. Machine-learning techniques enable computers to analyse vast data sets that would be far too large and complex for any conventional statistical tool to assess or human to comprehend (Raschka, 2015). The use of machine learning is ubiquitous in the modern world. For example, Amazon and Netflix tailor recommendations to individual customers based on their purchasing habits and those of millions of other customers; credit companies assess their ever-growing databases of transactions to identify patterns of fraudulent spending to freeze

accounts, and optical recognition tools continue to learn and identify handwriting (Coglianese and Lehr, 2016).

The use of machine-learning techniques has spread to the regulatory environment and promises substantial benefits (Yeung, 2016). With advances in machine-learning, it is now possible in principle to examine vast historic data sets and determine precisely which collection of metrics best prioritise inspections, excluding imperfect 'expert' judgement. In contrast with risk assessment tools that aggregate metrics selected *a priori*, machine learning is non-parametric; the algorithms "allow the data to dictate how information contained in input variables is put together to forecast the value of an output variable" (Coglianese and Lehr, 2016, p.7; Berk, 2008). With the algorithm objectively determining the optimal combination and weighting of metrics to predict outcomes, expert interpretation is unnecessary, often impossible given the size of the regulated sector and complexity of the model, and, as explored above, actively harmful to the accuracy of predictions. In data-driven approaches, risk assessment models will not only be more accurate than their predecessors developed by belief, intuition and consensus, they will not suffer from the biases that lead to poor expert judgement.

Purely data-driven approaches have been adopted by a number of government bodies. In 2004, the U.S. General Accountability Office identified over 50 federal agencies engaged in 'data mining' activities (General Accountability Office, 2004). The U.S. Environmental Protection Agency (EPA) has developed a machine-learning tool to prioritise a subset of the tens of thousands of new chemicals developed each year for comprehensive testing (Kavlock *et al.*, 2012). The U.S. Internal Revenue Service (IRS) uses machine-learning algorithms to prioritise tax collection from self-employed individuals and small business most at risk of not paying, and tax returns for review (DeBarr and Harwood, 2004; Martin and Stephenson, 2005). In the UK, the author has been involved in the development of machine learning prioritisation tools for Ofsted and CQC.

Further to removing the requirement for *a priori* metric selection and weighting, and problematic expert interpretation, purely data-driven approaches have a number of other advantages. *First,* eliminating expert interpretation and, usually, the more automated processes that accompany data-driven approaches both reduce costs. *Second*, the machine-learning algorithms can constantly learn and improve themselves, and be continuously monitored and updated with new data. *Third*, the risk assessments are empirically derived and hence provide a robust defence of prioritisation decisions in the face of political or media scrutiny.

These advanced data-driven approaches, however, also have their limitations. Depending on the machine-learning approach selected, prioritisation models can be incredibly complex which can

make it hard to justify why an individual regulatee has been prioritised for inspection. In practice, the more complex approaches, such as *support-vector machines*, are not used in regulatory environments where there is a requirement to fully understand the model (Schutt and O'Neil, 2013). Purely data-driven approaches also face difficulties trying to incorporate tacit knowledge that is not easily or readily quantified. Furthermore, the relationships identified between the data and outcomes can only be regarded as correlations, rather than causal inferences. As touched upon in section 3.5.1, if an algorithm tends to predict that larger schools are more likely to fail their Ofsted inspection than smaller schools, one cannot claim that increasing the size of a school will increase its probability of failing its next Ofsted inspection.

Data-driven approaches, therefore, use machine-learning techniques to objectively determine what collection and weighting of metrics best prioritise inspections. Such approaches are not reliant on the *a priori* selection of metrics based on intuition and consensus, but offer an evidence-based and objective model not hampered by flawed expert interpretation that can be continuously monitored and updated. More complex machine-learning models can, however, prove challenging to explain and, as with all quantitative models, will struggle to incorporate tacit information that cannot easily be quantified or categorised. Providing the most accurate and objective prioritisation models, and being best suited to continual monitoring, data-driven approaches best fit the stated goal of *Students at the Heart of the System* and offer the greatest promise in successfully introducing a risk-based approach to prioritising quality assurance reviews in higher education.

In summary, risk-based systems for prioritising inspections can be broadly grouped into three approaches. *First*, simple, rules-based methods place regulatees into prioritisation groups based on a small number of often contextual metrics selected *a priori*. *Second*, data-informed approaches present and aggregate numerous, wide-ranging metrics selected *a priori* to inform prioritisation decisions by one or more experts. *Third*, data-driven approaches use machine-learning techniques to determine the most accurate prioritisation model from the available data. It is this data-driven approach that best fits the stated goal of *Students at the Heart of the System* - namely to target QAA reviews via the objective assessment of a basket of data, monitored continually but at arm's length - and offers the greatest chance of success. There has been limited academic discussion of the risk assessment methods a risk-based approach to quality assurance in UK higher education should take. This literature is reviewed below in the context of the three broad categories of risk-based approaches discussed above.

3.6. Prioritising Quality Assurance Reviews in Higher Education

The purported benefits and challenges of a risk-based approach are well documented. The specific challenge of determining risk and prioritising regulatory activity as part of a risk-based approach, however, has received limited academic attention. The literature pertaining to quality assurance in higher education has not deviated from this pattern. Whilst there has been much discussion of the number and scope of regulatory bodies (BIS, 2015b, 2016; Brown and Bekhradnia, 2013; HEC, 2013; Universities UK, 2015), there is a notable lack of detail in the academic literature concerning *how* risk should be assessed. It is the specifics of a risk assessment as part of a risk-based approach to quality assurance in higher education that the final section of this chapter will focus on.

Rather than proposing an approach, Roger Brown, former Chief Executive of HEQC, argues against the use of performance indicators to determine risk, the cornerstone of *data-informed* and *data-driven* approaches, arguing "past experience can never be a reliable guide to future performance, especially when bearing in mind how quickly things can change" (McClaren and Brown, 2013, 42). Whilst it is true that past performance is not a perfect predictor of future performance, discounting it entirely as it is not perfect is excessive. Few realistically expect risk-based approaches to be 100% accurate, the very notion of a risk-based approach entails the acceptance of risk and impossibility of eliminating all harms (Demeritt *et al.*, 2015). The dismissal of past performance data on the grounds that it can never be a reliable guide of future performance is demonstrably wrong when considering the earlier studies assessing the performance of simple statistical models and expert judgement. Moreover, it is somewhat counterintuitive when considered in real world terms: it would be a brave patient who chooses the surgeon who has never had a successful operation over one that has never had a failure based on the assertion that past performance is no guarantee of future success. In the same paper, Anthony McClaren, Chief Executive of QAA until Summer 2015, proposed a combination of performance and contextual measures but offered no specific details as to what these measures may be or how they would be combined (McClaren and Brown, 2013).

King (2011b) has suggested approaches that would combine performance data with contextual information. Indicators of performance would be combined with measures of a provider's operating environment, internal governance arrangements, the nature of activities it undertakes, and the wider impact of governmental and regulatory policy. However, no proposal was offered for how the information should be combined, or whether the output from the risk assessment would be automatic, as is the case with most *rules-based* approaches, or determined by expert interpretation, as one would expect from a *data-informed* approach.

In 2013, the Higher Education Commission (HEC) published its "Regulating Higher Education" report which suggested a risk assessment method similar to a *data-informed* approach. The HEC advocated the limited use of past performance data combined with discretionary expert judgement, but were careful to note the need to avoid a formalised indicator set that one would expect of a *data-informed* approach. The use of a more formalised and scoped indicator set, the HEC argued, was not supported by evidence and would require "a highly sophisticated, but onerous, information strategy that would have led to sector reluctance to comply, as seen in Australia" (HEC, 2013, p.60). This argument overlooks the fact that there is also little evidence to support *any* method of risk-based assessment in a regulatory environment, and that the use of existing data would add nothing to the burden experienced by higher education providers.

Building on the work of the HEC, King (2014a) advocated a *data-informed* approach. King noted "the key to making a good regulatory judgement about risk and uncertainty … is not gathering masses of quantitative data but devising deliberative and other processes that help to weigh and judge data sensibly" (King, 2014b, p.4). Somewhat confusingly however, King also noted that the rise of 'big data' and the insights it can bring should also be included as, following a substantial increase in the information available concerning higher education provision, statistical correlations may exist which could forecast the probability of risks occurring and therefore trigger concern. King further noted that care should be taken however to understand the factors underlying any correlation and to ensure that the volume of available data does not overwhelm any model and that the increased collection and use of too much data in the pursuit of accurate predictions can see the 'signal' drowned out by the 'noise'. Whilst King is correct to urge caution over ensuring an understandable link between indicators and outcomes, the remaining 'big data' issues highlighted are precisely those the intrinsically linked machine-learning approaches were designed to solve.

The proposals by McClaren and Brown (2013), King (2011a) and the HEC (2013) demonstrate the underlying confusion in higher education research over what constitutes 'risk-based' regulation. The approaches they recommend, focusing on prioritisation using undefined risk measures and significant expert discretion, differ little from the subjective, implicit resource allocations that preceded development of a formal risk-based approach. It is unlikely such approaches would either reduce burden, reassure the public or be regarded as defensible by the QAA. In addition to lacking the explicit, objective risk assessment that defines a risk-based approach, the proposed approaches are lacking in detail over the cost and practicalities. What these approaches do provide is a demonstration of the lack of comprehensive understanding of risk-based approaches in higher education and the clear need for empirical evidence to underpin the debate.

The general approaches proposed in the limited literature favour the use of what could broadly be categorised as *data-informed* approaches utilising expert interpretation. No evidence, however, has been provided for the effectiveness of expert interpretation, merely theorised failings of the alternative, nor have any specific measures or means of combining them been proposed. What the evidence reviewed in the previous section has shown is that simple models comprising metrics selected by statistical processes will perform better than experts paying attention to metrics of their own choosing, or even when provided with the output from said model.

In summary, it is clear there is a range of opinions over how a risk-based approach to quality assurance in higher education should be operated. Although wide-ranging, these opinions are consistent in demonstrating confusion over statistical methods and what constitutes a risk-based approach: the solutions proposed are not objective, cost effective, burden reducing, robust or defensible. What the proposals do demonstrate is the need for an empirical assessment of available data to inform the use of risk-based approaches to quality assurance in higher education.

3.7. Conclusion

Risk-based approaches exploded in popularity at the start of the 21st century and continue to be a central tenet of the UK's regulatory landscape. The ostensible rationale – the quick and effective targeting of limited resource to where it will provide the greatest benefit whilst freeing high-quality providers from unnecessary regulation – is appealing to all parties. The literature however has identified numerous challenges to successfully operating a risk-based approach and realising its purported benefits. These include: political interference, immovable public perceptions, ill-defined or contradictory regulatory frameworks, the differing attitudes of regulatees, and resource limitations. All these challenges may however be redundant if a regulator cannot accurately determine risk as part of a burden reducing, cost effective approach. Despite this there have been no empirical studies – in higher education or otherwise – of whether this is possible.

The limited discussion of risk-based approaches to quality assurance in higher education has been lacking in technical detail or empirical underpinning. The proposed solutions have consistently demonstrated the confusion over what constitutes a risk-based approach and the statistical methods available to support them. This chapter has identified three broad approaches for assessing risk and prioritising regulatory activity which have developed as risk-based approaches,

and machine-learning techniques, have matured: *rules-based*, *data-informed* and *data-driven*. It is this *data-driven* approach that best fits the stated goal of *Students at the Heart of the System* and offers the greatest promise in successfully introducing a risk-based approach to prioritising quality assurance reviews in higher education (BIS, 2011). It is this approach therefore that this thesis focusses upon. The following chapter details the specific research questions and the data and methods which will be used to answer them.

4. Data and Methods

The purpose of this study is to provide an empirical analysis of whether the approach envisioned in *Students at the Heart of the System* is achievable: can the available data predict the outcome of QAA reviews, and hence prioritise them, as part of a risk-based approach to quality assurance? This chapter will detail the data available to answer this question, how it was prepared, and the chosen modelling and evaluation approaches.

4.1. Selecting the Data

The specific research questions and the way in which they are answered are necessarily dependent on the nature of the data available. Described below is how the dependent and independent variables were selected, what data was excluded and why, and the level at which the models were developed.

4.1.1. Dependent (Outcome) Variable

QAA's role is to assure quality. As noted in chapter two, this is defined by whether providers are meeting the 'expectations' detailed in the Quality Code (QAA, 2011b). Where providers are not meeting the expectations, QAA acts to ensure changes are made such that the expectations are met. It is where QAA finds issues, i.e. where expectations are not being met, that QAA will have impact. The dependent variable must therefore be the outcome of the QAA reviews. Using the outcome of inspections as the dependent variable in regulatory models is widely accepted practice and the approach adopted by *inter alia* Ofsted, the General Medical Council and the Food Standards Agency (Ofsted, 2015b; Lloyd-Bostock and Hutter, 2008; Food Standards Agency, 2016).

Some argue that QAA reviews do not get to the heart of quality (e.g. IUSSC, 2009a) and so the analysis should focus on predicting a more 'effective' measure of quality. Whether or not QAA reviews get to the heart of quality, they will have no real effect at providers that are meeting the expectations of the Quality Code, but that are not meeting some alternative definition of quality not used by QAA. Just as there would be little benefit in sending food standards inspectors to perfectly hygienic restaurants serving food that is not to everyone's taste, there is little benefit in QAA prioritising their reviews based on a measure of quality different from their own. The purpose of this thesis is to examine whether a risk-based approach can be used effectively to prioritise QAA reviews, and therefore it is the outcome of these reviews that are examined.

As noted in chapter two, in theory risk-based approaches focus on 'risk' defined as the product of the likelihood of an event occurring and its impact. In practice, however, risk-based quality regulators have shied away from incorporating such 'impact' measures as they have not wanted

to signal that some service providers are more worthy of their attention than others, or that any failure is acceptable (see for example CQC, 2013b; QAA, 2013a; Ofsted, 2015a; HEFCE, 2016c). Similarly, this analysis also focuses solely on the likelihood of an event occurring, i.e. of a provider receiving a negative QAA review, and not the perceived impact. This is for two reasons. *First*, it is not clear 'impact' would be used in the prioritisation of reviews, whereas it is certain the likelihood that a review will be 'unsatisfactory' would be. *Second,* how 'impact' should be measured is contestable and no measure of impact has been identified or agreed upon by actors in the higher education policy sphere. What is more, we know that if the likelihood of the review outcome cannot be predicted then, regardless of how impact is defined, a risk-based approach cannot work. Conversely, if the likelihood of an outcome can be predicted, a risk-based approach can work whether impact is defined as the number of students at a provider, the amount of taxpayer money at risk, the damage to the reputation of UK higher education, or as is most likely the case, not considered.

All electronically-available, complete QAA reviews were extracted from the QAA's past and current databases in late November 2014. A total of 2,847 reviews concerning 888 distinct providers, utilising 30 different review methods and dating from September 1999 to November 2014 were extracted. The reviews were then assessed to determine which methods were comparable with the current (2014) approach, i.e. conducted at provider-level and providing judgements on some or all of the current review questions:

- The setting and/or maintenance of academic standards
- The provision of teaching and learning opportunities
- The provision of information
- The enhancement of the quality of students' learning opportunities[3]

(QAA, 2013a)

This provisional list was then reviewed by the QAA and a final set of reviews comparable to the current approach, along with mappings for past questions and judgements to the current terminology, were agreed. The final data set comprised 853 reviews of 695 distinct providers, utilising 10 different review methods and dating from May 2007 to November 2014. This represents 62% of all reviews undertaken by the QAA during that time period. The majority of

[3] For HEIs the enhancement question was only introduced into the reviews considered in this study as part of the *Institutional Review* method in 2012/13. It was only introduced for FECs in Summer 2013 as part of the *Review of College Higher Education* method and for alternative providers in Summer 2014 as part of the *HER (Plus)* method. This is the reason for the low numbers of enhancement judgements seen in Table 4.1

those excluded were 'Developmental Engagement' reviews performed at subject level (QAA, 2005), or reviews with a narrow, specialist scope such as 'Early Years Professional Status Audits' or quality assurance reviews undertaken for the General Osteopathic Council (QAA and GOsC, 2011). The data was then cleaned with missing fields, such as the start date for the review, manually added where necessary, and mergers and name changes accounted for to ensure continuity where appropriate in the data.[4]

The QAA define a review as 'satisfactory' if all judgements were either 'Meets UK expectations' or 'Commended' (QAA, 2014d). Any review containing a judgement of 'Requires improvement to meet UK expectations' or 'Does not meet UK expectations' is classified as 'unsatisfactory'. Tables 4.1 and 4.2 below show the outcome of the 853 reviews by question-level and overall review-level judgements respectively.

	Academic Standards			Teaching & Learning			Information			Enhancement		
Question-Level Judgement	**HEI**	**FEC**	**Alt. Provider**	**HEI**	**FEC**	**Alt. Provider**	**HEI**	**FEC**	**Alt. Provider**	**HEI**	**FEC**	**Alt. Provider**
Commended	N/A	N/A	N/A	1	8	1	1	2	0	7	11	0
Meets	181	339	292	183	329	299	34	333	306	43	47	6
Req. improvement	10	5	21	7	3	19	1	4	1	0	13	3
Does not meet	0	2	11	0	6	10	0	7	22	0	2	0
Total	**191**	**346**	**324**	**191**	**346**	**329**	**36**	**346**	**329**	**50**	**73**	**9**

Table 4.1: The number of QAA reviews comparable to the current approach by question, judgement, outcome, and sector between May 2007 and November 2014.

	Overall		
Overall Review-Level Judgement	**HEI**	**FEC**	**Alt. Provider**
Satisfactory	178	320	286
Unsatisfactory	13	27	42
Total	**191**	**347**	**328**

Table 4.2: The number of QAA reviews comparable to the current approach by overall review-level judgement and sector.

As shown in Table 4.1, for HEIs, the number of 'Does not meet UK expectations' and 'Requires improvement to meet UK expectations' judgements is low in absolute terms. Indeed, no HEI has been judged 'Does not meet UK expectations' for any of the four individual questions assessed. Whilst this is good news for the sector, such low numbers are a cause for statistical concern. Developing a model based on too few outcomes can result in a model which is susceptible to

[4] Where a straightforward merger took place, for example with *University of Wales, Lampeter* and *Trinity University College* merging to form the *University of Wales, Trinity St David* in 2010, historic performance data was calculated using the aggregated data from predecessor bodies where possible.

'overfitting': a situation whereby the model predicts every sample perfectly having learnt not just the general patterns in the data but the unique and specific 'noise', the random statistical variance, of each occurrence (Babyak, 2004; Schutt and O'Neil, 2013; Kuhn and Johnson, 2013) (see section 4.3.1 and Appendix C for a fuller discussion of overfitting). To limit the effect of the low number of 'Requires improvement to meet UK expectations' judgements and there being no 'Does not meet UK expectations' judgements, the dependent variable was considered at overall review level, rather than question level, for HEIs. This resulted in final data set of 191 HEI reviews, 13 of which were classified as 'unsatisfactory'.

We are therefore looking to predict which HEIs will fail *any* of the four questions reviewed by the QAA. We are not looking to separately predict outcomes for *each* of the four questions reviewed by the QAA; to do so in a robust way with such low numbers is not possible. Moreover, as acknowledged by HEFCE, there would be little reduction in burden if shorter, more focused reviews looking at only a subset of the four review questions took place "as the preparations and documentation requirements remain the same" (2012b, 74).

For FECs, the numbers judged 'Does not meet UK expectations' or 'Requires improvement to meet UK expectations' at individual question level were also low. There were 347 FEC reviews in the data set of which, overall, 320 were 'satisfactory' and 27 were 'unsatisfactory'. Not all of these reviews could be used however. As discussed in greater detail in section 4.1.2.2 below, the two key datasets available which could provide metrics for model building relate to financial accounts and student characteristics. Without these two rich data sets there is very little data with which to predict the outcome of FEC reviews. The finance and student characteristics data sets are only available for the years 2007/08 and 2008/09 onwards respectively and both take considerable time to be published. Furthermore, to include trend analysis (change-over-time metrics) in order to look not just at current performance but also an FECs 'direction of travel' requires more than one year of data. Therefore, only reviews conducted after July 2011, once sufficient data were available, could be included in the analysis. The resulting data set, broken down by question and judgement, is shown below in Table 4.3:

FECs	Academic standards	Learning opportunities	Information	Enhancement	Total
Commended	N/A	8	2	11	21
Meets	159	147	155	47	508
Requires improvement	3	3	4	13	23
Does not meet	1	5	3	2	11
Total	163	163	164	73	727

Table 4.3: A breakdown of FEC reviews comparable to the current approach, and for which financial and student characteristics data was available, by question and judgement.

Again, low numbers and proportions of 'Does not meet UK expectations' and 'Requires improvement to meet UK expectations' judgements means that predicting the specific ordinal judgement for each question is not possible without a strong likelihood of overfitting the model. Thus, as with HEIs, the data was best considered at review level rather than question level. Again, this study therefore sought to predict which FECs would fail in *any* of the four questions reviewed by the QAA, not to separately predict outcomes for *each* of the four questions reviewed by the QAA. The reduced data set contained 143 'satisfactory' reviews and 21 'unsatisfactory' reviews from the period July 2011 to November 2014.

For alternative providers the picture was different. There have been a significantly greater number of reviews which can be included in the analysis and a greater number of incidents of providers receiving judgements of 'Does not meet UK expectations' or 'Requires improvement to meet UK expectations'. The result is that three of the four questions could in principle be analysed individually. Whether this is desirable or beneficial is another matter. *First*, it has already been shown that little resource can be saved, or burden reduced, by conducting reviews focused only on a subset of the QAA's four questions (HEFCE, 2012b). *Second*, the purpose of the models being developed is to prioritise reviews. It is far easier to develop and interpret models where the output is a predicted likelihood of a provider failing their review on any one question or multiple questions, rather than four distinct probabilities for each provider predicting the likelihood of failing each specific question.

First and foremost, this study therefore sought to predict which alternative providers would fail in *any* of the four questions reviewed by the QAA, not to separately predict outcomes for *each* of the four questions reviewed by the QAA. Subsequently, the *academic standards* question and the four possible ordinal judgements resulting from it were assessed to explore if accurate predictions could be made at question level. The *teaching and learning* and *information* questions, where there are fewer judgements in each of the four possible judgement categories, were then considered at a binary 'unsatisfactory' / 'satisfactory' level to determine if accurate predictions can be achieved at this level. The *enhancement* question was not analysed individually as only nine alternative provider reviews in the data set have assessed this question.

4.1.2. Independent (Predictor) Variables

In order to determine which metrics best predict the outcome of a QAA review, or failing that whether such predictions are impossible, it is necessary to consider as full a set of data as possible.

An initial review of the HE data landscape was undertaken, including documentation from the Joint Performance Indicator Working Group (JPIWG), Association of Colleges (AoC), BIS, HEFCE, HESA and the QAA. This initial review was then complemented by discussion with HESA, QAA, the QAA's external Research Advisory Group, and other relevant experts to establish what data was available.

The more comprehensive the data set is, the better the chances of developing an effective model to predict the outcome of a QAA review. However, common sense dictates that some metrics cannot feasibly have any direct causal impact on the likelihood of a provider having quality assurance issues, such as *the estimated percentage of staff/students who travel to work in single occupancy car journeys as their primary mode of travel* (HESA, 2016). Metrics lacking the slightest feasible link to quality and/or quality assurance were therefore discounted at this stage and not included in the analyses. Also discounted was information prohibitively resource intensive to obtain: not doing so would hinder the supposed 'efficient' rationale behind a risk-based approach (Hutter, 2005; Hampton, 2005). For example, background checks on each director or trustee of a provider may be indicative of risk but obtaining the data for each provider takes several person-hours; there is frequent director turnover; and as such sourcing and regularly updating the data is not feasible. The data available for each provider type are discussed in turn below.

4.1.2.1. HEIs

As a long established, quasi-public sector domain traditionally in receipt of large amounts of public money that carries with it substantial reporting requirements, the HEI sector is far more data rich than the FEC or alternative provider sector. Once all the suitable data had been identified it was sourced from the *Higher Education Information Database for Institutions* (HEIDI) or via the QAA for the academic years 2003/04 – 2012/13 where available. The data comprised:

- Applications data – this includes applications made via the Graduate Teacher Training Registry, University and College Admission Service (UCAS), Conservatoires UK Admissions Service, and Nursing and Midwifery Admissions Service.
- Destinations of leavers from HE (DLHE) Survey - the DLHE survey asks leavers from higher education what they are doing six months after graduation.
- Research Statistics – this includes market share of research grants, contracts income, research staff and research council research studentships by institution.
- Unit Expenditure Statistics – the expenditure of departments, academic services and administrative services.

- HESA Performance Indicators – derived metrics covering a number of topics including leavers, research, participation of under-represented groups in HE, and continuation rates.
- Staffing metrics – staffing details by academic employment function, mode of employment, nationality, source of salary, and terms of employment.
- Student metrics – student details by mode of study, degree classification, age, domicile, gender, and nationality.
- Staff Student Ratios
- Finance Indicators – both key financial indicators (KFIs) and full accounts information.
- Aggregate Offshore Record – details of students studying overseas for a UK HE qualification
- National Student Survey - a high-profile survey aimed at mainly final-year undergraduates which gathers opinions relating to six aspects of the learning experience, including one question about overall student satisfaction.
- Previous Quality Assurance Reviews – the outcome of previous, comparable QAA reviews.
- *QAA Concerns* – the QAA has a "concerns" procedure for investigating systematic issues relating to academic standards, learning opportunities and the provision of information that individuals do not feel have been satisfactorily addressed by the higher education provider in the first instance.

In total there were 751 HEI metrics prior to any variants being calculated. For a full list of HEI metrics used in this thesis see chapter five, Appendix E.

4.1.2.2. FECs

The data available to potentially prioritise QAA reviews of higher education in further education colleges (HE in FE) as part of a cost-effective, risk-based approach is a fraction of that available for HEIs. This is in part due to there being less mandated and centrally controlled data to capture: applications are often not made through UCAS, and the majority of HE in FE providers do not undertake research or have an overseas campus for example. The main reason for the lack of available data however is the diverse nature of delivery in the HE in FE sector and the fractured reporting that results. Provision can be financed directly or indirectly by funding councils or via bodies such as the Skills Funding Agency (SFA). Different funding channels carry with them different data definitions and reporting requirements, which makes reporting on HE in FE a

challenge. Learners in England receiving directly-funded, prescribed provision[5] must be recorded on the Individual Learner's Record (ILR) and the Higher Education in Further Education Students (HEIFES) early statistical return submitted to HEFCE. Learners receiving indirectly-funded, 'prescribed' provision are recorded by the franchising institution (typically a university) and included in their HESA returns along with their direct-provision students and in the Higher Education Student Early Statistics (HESES) return. FECs should not include details of indirectly-funded students in their ILR returns and should instead report the number of such students separately; however, instructions on the matter have previously been unclear and not adhered to leading to double counting and significant concerns over the quality of student-level data (Storan and Hudson, 2015). The challenges of combining data sources with different data definitions and multiple reporting issues has led to warnings concerning the accuracy and interpretation of data in previous HE in FE research (Clark, 2002; Parry and Thompson, 2002; Tait *et al.*, 2008).

Once the potential metrics had been identified the data was sourced from QAA, SFA, Ofsted, and HEFCE. This resulted in five key data sets summarised below:

- *QAA Concerns*
- Previous Quality Assurance Reviews
- Ofsted Rating – the published Ofsted rating (which relates to non-HE provision) for the FEC at the time of the QAA review.
- Student Characteristics Data – specifically for this study HEFCE have supplied the student characteristics data for HE students studying at FECs, whether that provision is franchised or not, by linking the ILR and HESA data set. These data break down student numbers by gender, first year/non first year status, age group, domicile, mode of study, ethnicity, and level of study[6].
- Financial Accounts – the financial accounts of all FECs containing in excess of 350 metrics for each FEC are quality assured and published by the SFA.

[5] When a college becomes 'directly funded', it receives funding to support a range of HE activities. Colleges sign a 'funding agreement' with the funder, which sets out 'conditions of grant'. These conditions mean the college must comply with requirements, for which we hold the college directly responsible. Direct funding also means that students at the college can access government loans and grants.

'Prescribed' courses are courses that lead to qualifications set out in the Further and Higher Education Act (1992). These are: higher degrees, postgraduate diplomas, postgraduate initial teacher training qualifications (such as PGCE), first degree (including foundation degree, BSc, BA, Bed), foundation degree bridging course, HND, DipHE, HNC, 120-credit point Diploma in Education and Teaching (DET), and CertEd (HEFCE, 2016a) .

[6] Whilst these data are not normally published at this level, they would be easily available for any agency-based review.

In total there were 181 FEC metrics prior to any variants being calculated. For a full list of FEC metrics used in this thesis see chapter six, Appendix F.

4.1.2.3. Alternative Providers

There are fewer data available concerning alternative providers than for HEIs of FECs. As discussed in chapter two, alternative providers interact with QAA either when applying for course specific designation (RCSD), or when applying for 'Tier 4 Sponsor Status' (REO). Due to contractual reasons, the detailed FSMG checks, which are conducted by a third party, cannot be systematically shared with QAA or anyone else and therefore could not form part of this analysis.

As alternative providers do not receive direct funding from HEFCE or its equivalents there have previously been no other reporting requirements placed on them, other than those which apply to any other registered company. Alternative providers who receive course designation have been required to report a student record data set to HESA from 2014/15 onwards; however, these data come too late for this study, are of unknown quality, and account for less than a quarter of QAA's alternative provider reviews in the data set.

The only data that was centrally available and covered the majority of the alternative providers were:

- *QAA Concerns*
- Previous Quality Assurance Reviews
- Companies House records - In order to obtain further relevant information, the two most recent sets of financial accounts available prior to each provider's review, where filed[7], were purchased from Companies House and the elements common to all accounts, the balance sheets, were transcribed.

In total there were 38 metrics concerning alternative providers prior to any variants being calculated. For a full list of alternative provider metrics used in this thesis see chapter seven, Appendix G.

[7] Companies are given 21 months after first registering with Companies House to file their first year's accounts and nine months following the end of their financial year to file subsequent accounts. Nine alternative providers were so recently established that they had filed no accounts, eight providers were also specialist institutions – for example charitable religious organisations – for which company accounts could not be found.

4.1.3. Linkage Between the Dependent and Independent Variables

A data-driven approach, as explored in this thesis, automatically derives the best possible model for predicting the outcome of QAA reviews based on the available data. The *a priori* selection of metrics believed to be useful in predicting the outcome of reviews is not required. What is required *a priori,* however, is the decision on how the data set(s) should be constituted for the machine-learning algorithm to determine the optimal model(s).

Machine learning approaches cannot consider metrics that cannot be complete for all cases in the data set. For example, a machine learning approach could not predict the price farm animals would fetch at auction using a measure of the percentage of their milk that was drinkable as male farm animals cannot produce milk. For male animals the metric is nonsensical. This leaves the modeller with two options. They can develop a single model for all animals and not consider the metric regarding the quality of milk. This would result in one model, which would be simpler to work with than separate models for male and female animals, but may represent a loss of useful information. Alternatively, the modeller could split the data set and develop one model for female animals, which considers the quality of their milk, and a separate model for male animals which does not. These two models would be more effort to work with than a single, non sex-specific model, but the individual models for male and female animals could be no-less and more accurate respectively. This consideration impacts on the study in two ways: the exclusion of specific provider types from different countries within the UK, and the use of provider-type specific models.

4.1.3.1. Excluding nation and provider-type specific reviews

QAA is a UK-wide body (with QAA Scotland an independent organisation within QAA) responsible for ensuring quality and standards of all higher education delivered by UK establishments, whether in the UK or overseas. However, higher education is a devolved matter in the UK and the four national funding councils, HEFCE, HEFCW, SFC and DELNI reflect this. As a consequence, the complexities of different national legislative frameworks and goals have led the QAA to adopt different review approaches for some or all of the constituent nations for each provider type. It also means that different data sets are available for different nations. As there is no perfect alignment between national / provider type dependent and independent variables a decision has to be made about what should be included in the final data set. For example, with no student data available for the eight HE in FE providers in Wales, a choice had to be made between developing a model including these eight providers but necessarily excluding student and finance data for all HE in FE providers, or a model which did not cover the eight Welsh HE in FE providers but could

make use of the student and finance data. Given the importance of these data sets and the small percentage of HE in FE providers that were Welsh, the decision was made to exclude the Welsh HE in FE providers.

The national / provider type reviews that were excluded from the final data set, and the reason for their exclusion, are detailed below in Table 4.4:

	HEIs	FECs	Alternative Providers
Wales	Included	Excluded – student-level or finance data was not available. There are 8 HE in FE providers in Wales	Included
England	Included	Included	Included
Scotland	Excluded as the component questions of the *Enhancement Led Institutional Review* were incompatible with all other current review methods. There are 19 HEIs in Scotland	Not applicable – no Scottish FECs provided HE at the point the review data was collected.	Included
Northern Ireland	Included	Excluded – student-level or finance data was not available. There are 6 HE in FE providers in Northern Ireland	Included

Table 4.4: A breakdown of which provider types are included in each analysis by country.

In practice, the independent nature of QAA Scotland means that their activities would have been separate from the rest of the QAA and therefore the exclusion of Scottish HEIs does not significantly compromise any risk-based model.

4.1.3.2. Focusing on individual sectors

As described above, there is substantial variation in the volume and nature of metrics available to predict the dependent variable – the outcome of QAA reviews - in each sector. It is therefore preferable to focus on each provider type – HEIs, FECs and alternative providers – individually and make full use of the data available. The output from these three models could be combined for practical purposes should QAA require predictions for all providers to be considered together.

This approach also highlights which metrics that are not currently collected for a given sector may be worth investing in or not. For example, if a set of metrics were able to accurately predict the likelihood of HEIs receiving an 'unsatisfactory' QAA review and they were not available for FECs or alternative providers, then this study can evidence that these metrics should be prioritised for development beyond HEIs. Conversely, if despite the breadth and depth of HEI-related information available, no metrics could have predicted the outcome of past HEI reviews then this

may suggest the implementation of a data-driven, risk-based approach will face substantial difficulties even with the imposition of additional data collections for FECs and alternative providers.

Focussing on individual sectors therefore makes the best use of the sector-specific data and highlights whether the collection of such data in other sectors would prove worthwhile.

4.1.4. Summary

Due to the low number of 'Does not meet UK expectations' and 'Requires improvement to meet UK expectations' judgements, the greater ease of modelling and interpretation, and the fact that the resource required to review a provider in relation to one question or all four varies little, it is the overall outcome of the review, rather than the specific question-level outcomes, that were predicted. There is significant variation in the volumes of data available for each of the three provider types – HEIs, FECs and alternative providers – and, in order to make the most of this data, a separate model was developed for each sector. Due to differences in the delivery of higher education provision, data collections, and QAA review methods in each of the four nations of the UK, a subset of reviews was excluded from the overall data set.

4.2. Data Preparation

With the dependent and potential independent variables determined, the next stage for each sector was to prepare the data set for modelling. This required a number of processes as shown below in Figure 4.1:

1. Change-over-time variants for each metric were calculated.
2. Each review was matched with the most up-to-date version of each metric prior to the review.
3. The data set was then assessed to determine which reviews and metrics needed to be removed due to lack of coverage or their anomalous nature.
4. The remaining metrics with missing data were assessed.
5. Metrics containing missing data then either had missing values imputed (statistically estimated) or were removed (see section 4.3.4 and Technical Appendix D for a full explanation of imputation).
6. If appropriate, data was then also standardised or benchmarked.

7. Non-variant and highly correlated metrics are removed.

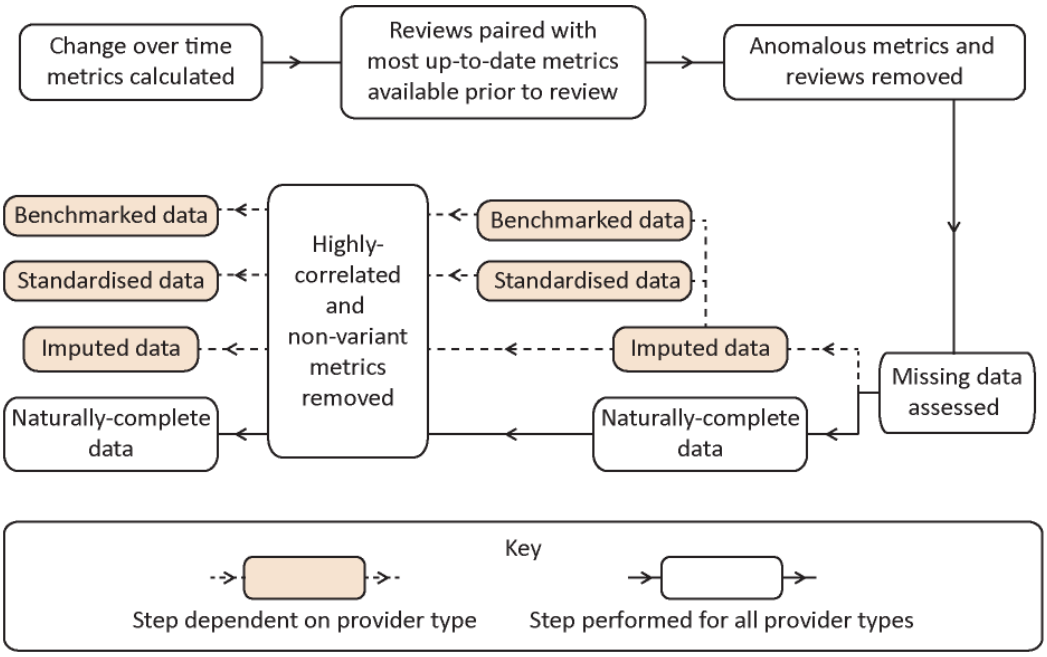

Figure 4.1: The ordered stages of preparing the data for statistical modelling.

Each of these stages is discussed in turn below followed by the details of the application of the whole process to each of the three provider types.

4.2.1. Change-Over-Time Metrics

Four change-over-time values were calculated for each applicable metric; these were the one and two-year absolute and percentage changes. An example of these calculations is shown below for a four-year period for an HEI's National Student Survey 'Overall Satisfaction' score:

	National Student Survey Overall Satisfaction Score by Academic Year			
Metric Code	2004/05	2005/06	2007/08	2008/09
Standard metric (Abs)	82.28	84.88	84.33	81.87
One-year change in absolute value (Ca1)	N/A	84.88 – 82.28 = 2.60	84.33 – 84.88 = -0.55	81.87 - 84.33 = -2.46
One-year change in percentage value (Cp1)	N/A	$\frac{(84.88-82.28)}{82.28}=0.032$	$\frac{(84.33-84.88)}{84.88}=-0.0065$	$\frac{(81.87-84.33)}{84.33}=-0.029$
Two-year change in absolute value (Ca2)	N/A	N/A	84.33 – 82.28 = 2.05	81.87-84.88 = -3.01
Two-year change in percentage value (Cp2)	N/A	N/A	$\frac{(84.33-82.28)}{82.28}=0.025$	$\frac{(81.87-84.88)}{84.88}=-0.035$

Table 4.5: Example calculations of one and two-year absolute and percentage change-over-time metrics.

For the first year of metric data no change-over-time variants could be calculated as no prior data was available. For the second year of metric data the one-year change-over-time metrics could be calculated but not the two-year change-over-time metrics. For this reason, where possible, two years' worth of data was collected prior to the first review for each provider type.

For all three provider types change-over-time variants were calculated for each metric where appropriate. Those metrics where change-over-time variants were *not* deemed appropriate were those that related specifically to previous reviews, for example *the number of prior reviews which were 'unsatisfactory'*. This is unlikely to be different one or two years prior to a review given the long spanning, cyclical nature of reviews. Moreover, amending the metric to look at the change since last review made little sense as it was often the case that providers had only had one prior review if any at all.

4.2.2. Matching Metrics and Reviews

Once the change-over-time metrics had been calculated all the data was uploaded into a Microsoft SQL Server database. A SQL query was then run to match each review with the most up-to-date version of each metric that was published 28 days prior to a review being undertaken. The requirement for data to have been published 28 days prior to the review is to allow a realistic amount of time for the QAA to source the data, format it, incorporate it into their predictive model and act on the outputs.

4.2.3. Removing Anomalous Reviews and Metrics

The first stage in dealing with the missing data was to determine if there were any providers which were missing data for a significant number of metrics that were complete for nearly all other providers. If so, theses anomalous providers were removed. For example, the self-financing Ashridge was removed because, although it was classified as a 'University/HEI' by QAA and subject to an Institutional Audit, it offers only postgraduate 'executive education' and does not submit data to HESA.

4.2.4. Missing Data Assessed and Addressed

At this stage the three data sets, one for each provider type, were large but contained a number of instances of missing data. For a metric to be useful it must have a value associated with each review in the final data set. Within the data collections available for this study there are two key reasons data can be missing:

- It may be 'structurally' missing, i.e. missing because it cannot exist. For example, the proportion of an HEI's postgraduate students who complete their studies cannot be captured for a HEI with no postgraduate study (you cannot divide by zero).
- Alternatively, it may be missing due to chance factors. An HEI may have filed their data return incorrectly leading to that information not being available for one year. Similarly, a chance flood of a records office could have destroyed the data for one year.

Where we do have missing data it can be dealt with in one of three ways: the metric can be removed entirely from the data set and the affected reviews retained; the review(s) affected by the missing data can be removed from the data set and the metric retained; or, if appropriate, the metric can be imputed to provide a statistical estimation of the missing value and both the metric and the review(s) can be retained.

The default approach for a statistical model is to remove those cases, (i.e. in this study the QAA reviews), which have any missing values for any of the metrics and this is one of the approaches adopted for all three provider types. However, it is argued that this approach may lead to estimates which are themselves biased, inaccurate, or both (Harrell, 2001; McKnight *et al.*, 2007; Van Buuren, 2012). Moreover, removing those QAA reviews which are affected by missing values on a metric, even if it's only a very small percentage, and retaining the metric in the model will lead to increased standard errors, widened confidence-intervals, and a decrease in the power of goodness-of-fit tests (Donner, 1982). The latter is of special concern for this data set where the number of reviews with an 'unsatisfactory' outcome is low. As an alternative, values 'missing at random' can be imputed – statistically estimated - using a number of techniques (Harrell, 2001; Van Buuren, 2012). For further details on the statistical imputation methods used in this thesis see Appendix D at the end of this chapter.

The first step when dealing with missing data therefore is to assess the impact on the overall data set of removing specific metrics and reviews which contain missing values and to determine whether those missing values can be imputed. Those metrics which contain structurally-missing data – that is data that is missing because it cannot exist – must be excluded regardless of whether imputation is to be used. The next step is to determine whether to impute the data. To many non-statisticians this approach may not seem acceptable: no matter how advanced the imputation method, data is still being invented. Therefore, where imputation was used, two data sets were created for the analysis, one using only naturally-complete metrics and a second including metrics with imputed values too.

With the structurally-missing data removed the data can then be imputed. For the naturally-complete data set not using imputation however, metrics and/or reviews with missing values need to be removed. Clearly where there is only one review with values missing for hundreds of metrics, it would be preferable to sacrifice that review to retain the hundreds of metrics relevant to all other reviews, rather than sacrifice all those potential predictors. Alternatively, if a metric is missing values for hundreds of reviews, the preference would be to sacrifice the metric in order to maintain the reviews. Deciding which to sacrifice is a matter of judgement and must consider the relative importance of the metrics and reviews affected.

4.2.5. Standardising or Benchmarking Metrics

In addition to considering a provider's absolute performance and their 'direction of travel' as determined by the change-over-time metrics, it may be beneficial to standardise the data to account for sector-wide changes in performance, or focus on assessing performance only in relation to similar providers.

The NSS provides a useful example to demonstrate the potential benefit of in-year standardisation. The average provider's NSS overall satisfaction score was 81.89% in 2009/10 and 84.76% in 2011/12. With a standardised metric, a provider performing exactly in line with the average in 2009/10, i.e. with a metric value of 81.89%, would have the same metric value as a provider performing exactly in line with average in 2011/12, i.e. with a metric value of 84.76%. The standardised metric was calculated using the average metric score and standard deviation for each academic year:

$$\frac{\text{Provider metric score} - \text{ average metric score for that year}}{\text{Metric standard deviation for that year}}$$

In the example above both providers would have a standardised score of 0 (their performance is exactly equal to the average for that year and so the numerator will be zero in both cases) whereas, with the normal, non-standardised metric, the provider reviewed in 2011/12 would be seen as performing better than the provider reviewed in 2009/10 by virtue of its higher absolute score. For the provider types where standardisation was deemed appropriate, only the metrics comparing a provider's annual performance to others were standardised. It does not make sense to standardise metrics relating to how a provider has performed in isolation, for example a provider's previous review outcome, nor would it make sense to attempt to standardise this by year.

Data on the proportion of students who successfully graduate provides a useful example to demonstrate the potential benefit of benchmarking. A Russell Group university will have a student

population that is, on average, academically higher-achieving than former newer institutions with a different mission. Comparing one's performance against the other may be less informative than comparing, for example, Russell Group universities against each other. If one Russell Group university is performing far worse than the others, but marginally better than a set of newer institutions, the fact that it is not performing as it could be with the resources it has available may be more telling than the lesser performance, in absolute terms, of the newer institutions with lesser resources. The benchmarked metric was calculated as:

$$\frac{\text{Provider metric score} - \text{average metric score for the provider's benchmark group}}{\text{Metric standard deviation for the provider's benchmark group}}$$

Benchmarking was only considered appropriate for HEIs where the provider missions vary significantly and identifiable groupings of providers exist. The benchmarking groups used were those identified by Wolf (2015): 'Russell Group', 'Other Old', 'Former Polytechnic', and 'Other New'. Not every HEI fit readily into one of these four categories; conservatoires, for example, are often well-established but are not comparable to the majority of 'Other Old' universities which dwarf them in terms of student numbers and diversity of subjects. For that reason, not every HEI was assigned a benchmarking group and 38 reviews were excluded from the analysis. This reduction in the number of reviews did however increase the number of complete, non-correlated metrics available (see 4.2.8.1 below for further details on metric numbers for the HEI analyses).

4.2.6. Non-Variant and Highly-Correlated Metrics

Despite having no missing values, some metrics can still be of little or no value when developing a predictive model. Indeed, the presence of redundant metrics can hamper model development unnecessarily by requiring additional computational effort and increasing the probability of a chance, meaningless relationship being identified resulting in a misleading model (Zhao and Cen, 2013).

To identify and process these metrics the data was read into the R statistical program. The data set was examined to identify metrics that had no variance, i.e. metrics that had the same value prior to every review, and therefore did not discriminate between providers in any way. An example of this would be a metric of whether or not HEIs hold taught degree awarding powers. All do and knowing this does not help predict which HEIs will be judged 'unsatisfactory'. Non-variant metrics were therefore removed at this stage. There were a number of instances of near-zero variance metrics, that is metrics for which values were near, but not fully, uniform across each provider, such as the proportion of students aged 17 for HEIs. These metrics were retained at this stage as there was no guarantee that the few providers that differ from the rest on the

measure in questions were not the same ones who received 'unsatisfactory' judgements. The data set was then assessed for highly correlated metrics. The corrplot package in R (Wei, 2013) allows for an algorithm to be run such that:

1. The correlation matrix for all predictors is calculated.
2. The predictors with the greatest pairwise correlation above a defined cut-off point are selected. For this study a high cut-off of 0.9 was used.
3. The average correlation between each of these selected predictors and all others in the data set is calculated.
4. The variable from the selected pair with the largest average correlation across all predictors is removed.
5. Repeat until no predictors with a pairwise correlation score above the defined cut-off.

To be clear this only represents a winnowing down of a cluster of highly correlated metrics such as *the number of full-time equivalent (FTE) staff from the EU* and *the number of full-person equivalent (FPE) staff from the EU*, both of which appear in the initial data set. Not all highly-correlated metrics are removed; the metric from each highly-correlated pair that is most unique amongst the wider data set remains.

4.2.7. Summary

In summary, to prepare the data for statistical modelling change-over-time metrics were calculated; each review was mapped with the most up-to-date version of each metric prior to the review; anomalous reviews and metrics were removed; the remaining metrics were assessed to determine whether values were 'structurally missing'; if appropriate a data set with missing values imputed was created along with a data set where all metrics with missing values were removed; if appropriate an additional data set containing standardised and benchmarked variants of relevant metrics was also created; and finally non-variant and redundant highly-correlated metrics were then removed from each data set.

4.2.8. Provider-Type Specific Data Preparation

Detailed below is the application of each of these steps to the three provider-type specific data sets.

4.2.8.1. HEIs

The HEI data set initially contained 751 metrics (for a full list see chapter five, Appendix E) which increased to 3,698 metrics with the addition of one and two-year absolute and percentage change metric variants. The most up-to-date version of each metric available 28 days prior to the start of the review was then matched with each of the 191 reviews available. The data set was then inspected and seven reviews were removed for their anomalous nature including 'Richmond, The American International University in London' which is a private provider offering US degrees but also has the option to award UK degrees through a validation arrangement with the Open University. This provider was classed as a 'University/HEI' by QAA and underwent an 'Institutional Audit' but is not directly funded, does not hold degree-awarding powers and does not submit data to HESA. Furthermore, 59 metrics were removed as unsuitable at this early stage due to having data for only a small fraction of the reviews. The cleaned data set was then separated into two. The first data set concerned all HEIs for which 2,009 of the metrics were deemed to contain 'structurally missing' data and were removed leaving 1,690 metrics. The majority of the metrics that were removed were proportions or percentages with possible, and often frequent, zero dominators such as *the proportion of UK-domiciled Postgraduate leavers who obtained qualifications through part-time study and entered further study (including those that are working & studying)*. The second data set concerned only those HEIs with an identifiable and satisfactory benchmark grouping, 38 reviews were removed but as a result there were a greater number of metrics not affected by missing data.

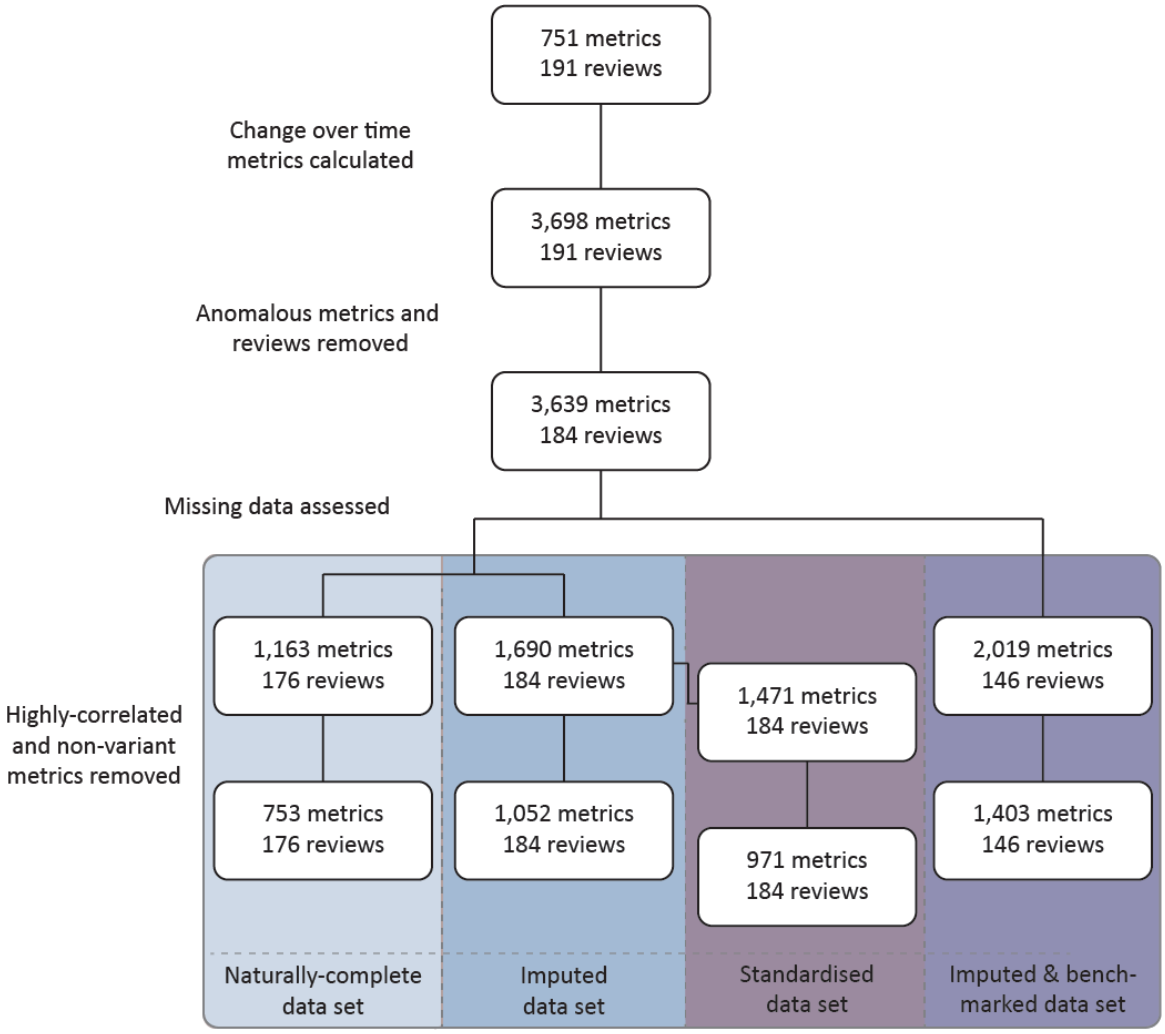

Figure 4.2: A summary of the data preparation process for HEIs.

At this point three data sets were formed from the non-benchmarked, 184 review, 1,690 metric data set. The first data set contained only naturally-complete metrics - that is metrics with no missing or imputed values. To begin with each metric and review contained in the data set was assessed to determine the number of missing values. The optimal balance of maintaining a large and broad range of metrics was obtained by removing eight reviews of six specialist institutions such as London School of Hygiene and Tropical Medicine which only offer postgraduate degrees. None of the reviews removed from the data set were 'unsatisfactory', the outcome of the reviews that we are most interested in and less common than the 'satisfactory' review outcome. The result was that only 527 metrics then needed to be removed to produce a naturally-complete data set.

For the second data set, all reviews were retained and all remaining metrics with missing values were imputed. None of the imputed metrics had coverage below 85 per cent. That high value meant that the imputation is more likely to be accurate and the analysis robust.

The third data set was the standardised data set deemed worth exploring for HEIs given the seven-year time span between the earliest and latest reviews in the data set and the changes to the sector that have occurred in this time. This data set was created by standardising the absolute metrics in the imputed data set and then recalculating the absolute change-over-time metrics based on the newly standardised metrics. The percentage change-over-time variants were not calculated as, once the data had been standardised, there is little difference between absolute and percentage changes in metric values. The resulting data set contained 1,471 metrics relating to 184 reviews. Finally, the highly-correlated and invariant metrics were removed leaving 753, 1,052, and 971 metrics in the naturally-complete, imputed and standardised data sets respectively.

4.2.8.2. FECs

The first step was to calculate one-year absolute and percentage change-over-time variants for the 181 metrics where this was appropriate. The two-year change-over-time variants were not calculated as this would have required sacrificing a further year of reviews which was deemed unacceptable (see section 4.1.1). Calculating the one-year change-over-time variants increased the number of metrics from 181 to 528. Next, eight anomalous reviews, all of which had a 'satisfactory' outcome, were removed from the data set leaving 155 reviews. Examining the data at this stage there was a clear division between metrics for which data was available for all providers and metrics for which most of the data was consistently missing for a significant number of providers. The FEC sector has far fewer metrics available than for HEIs but those that do exist are key metrics whose completion rate is high. The number of metrics available was therefore far smaller but there were far fewer cases of missing data. Those metrics that were missing values were doing so for structural reasons and imputing the data would provide no benefit. There was therefore no imputed data set for FECs. The data was then assessed to determine which metrics and/or reviews should be removed to create the naturally-complete data set. The only metric retained at the expense of reviews was the FEC's Ofsted rating at the time of their QAA review. This was because there was such a clear rationale for Ofsted inspection ratings to predict the outcome of QAA reviews: both are centred on reviews of quality of the FEC. The resulting data set contained 338 metrics relating to 131 reviews.

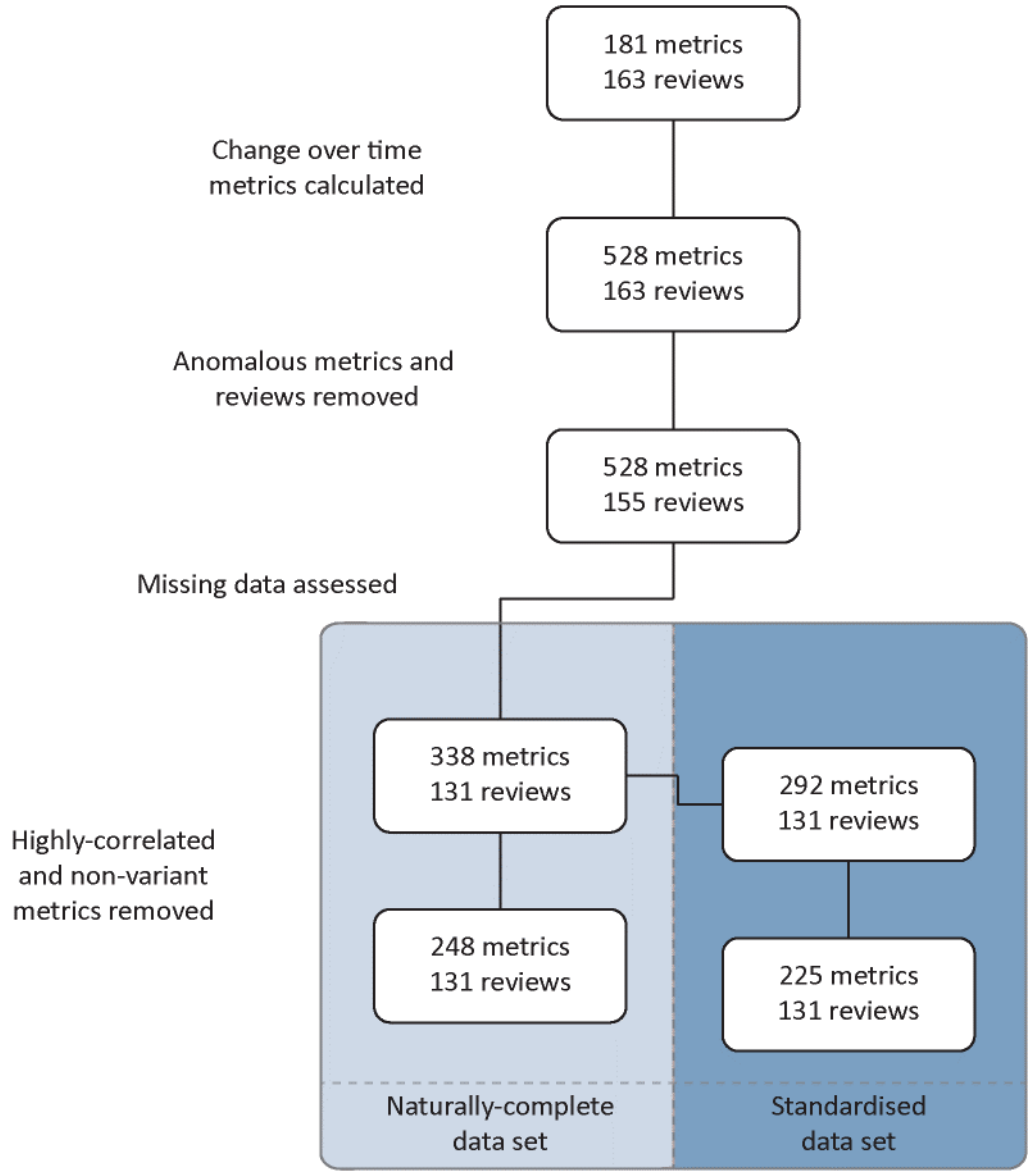

Figure 4.3: A summary of the data preparation process for FECs.

At this stage a second data set was created by standardising the absolute metrics in the cleaned data set and then recalculating the absolute change-over-time metrics based on the newly standardised metrics. This data set included 131 reviews and 292 metrics; 46 fewer than the naturally-complete data set due to redundant percentage change-over-time metrics not being calculated as these added no value to standardised metrics. Finally, the highly-correlated and non-variant metrics were removed leaving 248 and 225 metrics in the naturally-complete and standardised data sets respectively. The data was not benchmarked as no clear, meaningful benchmarking groups existed for HE in FE providers.

4.2.8.3. Alternative Providers

For alternative providers the first step was to calculate the change-over-time variant of each financial metric. More than a year can elapse between sets of accounts being filed with Companies House, especially for smaller providers with reduced reporting requirements. Accordingly, the change-over-time metrics were calculated as the change from the previous set of published accounts, rather than the change from the latest accounts available exactly one year prior. As with HEIs and FECs, no change-over-time variants were calculated for *QAA Concerns* and previous review outcome metrics as this added no value: existing metrics detailed all the historic review outcomes and *QAA Concerns.* The calculation of the change-over-time variants increased the number of metrics from 38 to 52. No reviews or metrics were deemed anomalous and in need of removal; however, financial accounts were unavailable for 23 reviews concerning 22 providers and these were removed. All but one these reviews had a 'satisfactory' outcome.

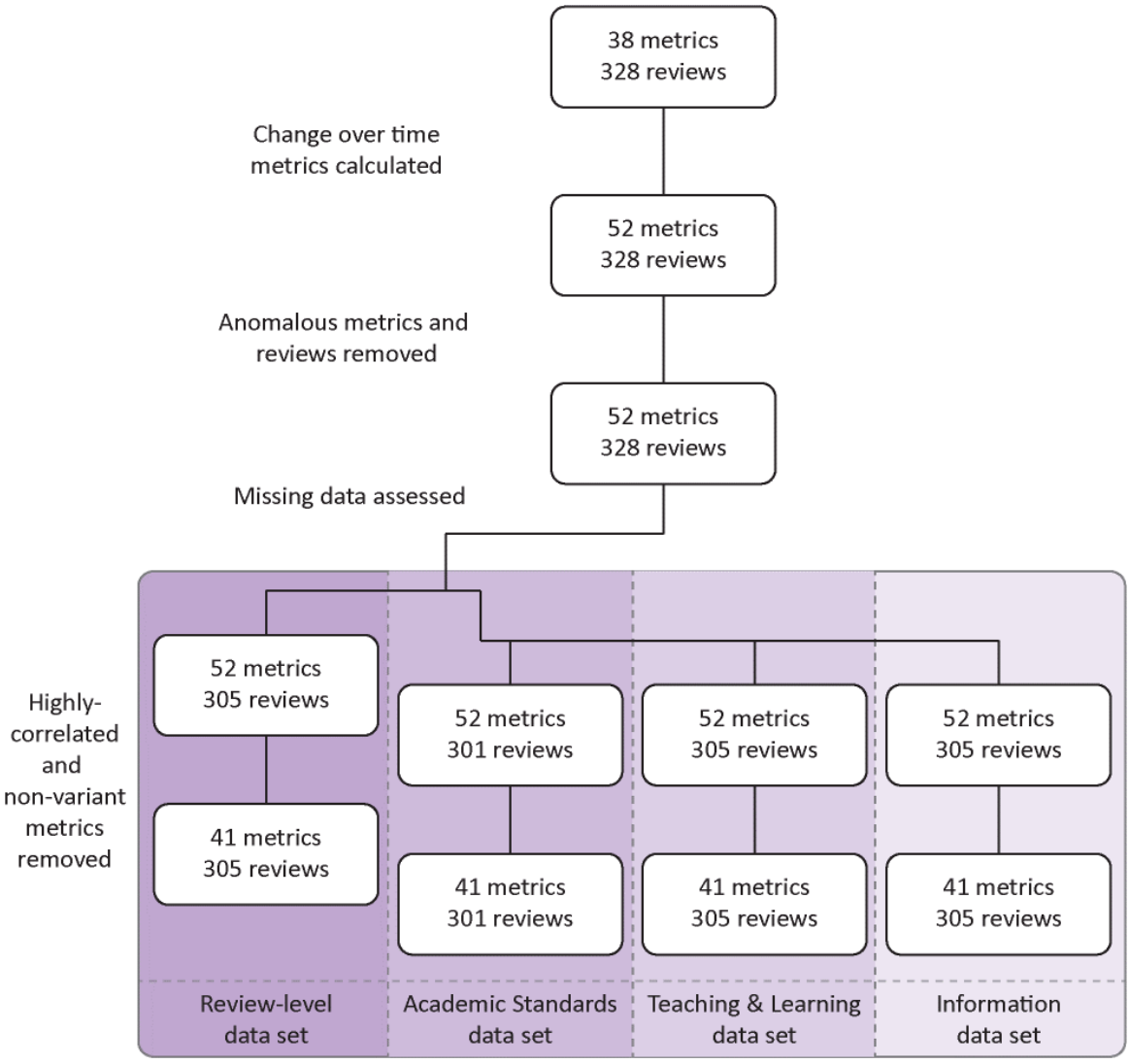

Figure 4.4: A summary of the data preparation process for alternative providers.

As little other information was available for these 22 providers prior to their reviews, there was little chance of imputing accurate and informed values for the missing data and therefore imputation was not used. Spanning less than three years and with a limited number of metrics suitable for standardisation it was also judged that there would be no benefit in creating a data set containing standardised versions of the metrics. Benchmarking was also considered inappropriate as no clear, meaningful benchmarking groups existed for alternative providers.

Four data sets were then created containing the appropriate dependent variable: the overall review outcome, the ordinal judgement (does not meet, requires improvement to meet, or meets UK expectations) relating to academic standards, and the binary judgement (satisfactory or unsatisfactory) relating to teaching and learning and the provision of information. In each case the highly-correlated and non-variant metrics were removed leaving 41 metrics in the data set.

4.2.9. Summary

To prepare the data for statistical modelling: change-over-time metrics were calculated, each review was mapped with the most up-to-date version of each metric prior to the review, anomalous reviews and metrics were removed, the remaining metrics are assessed to determine whether values were 'structurally missing', if appropriate a data set with missing values imputed was created along with a data set where all metrics with missing values were removed, if appropriate additional data sets containing standardised and benchmarked variants of relevant metrics were also created, and finally non-variant and redundant highly-correlated metrics were then removed from each data set.

Four data sets resulted for HEIs: one naturally-complete data set containing only metrics and reviews with no missing values; a data set containing eight more reviews and 527 more metrics which had their missing values imputed; a third data set that contained metrics that were standardised by academic year; and a final data set that contained metrics benchmarked by HEI type. For FECs, neither imputation nor benchmarking were appropriate but standardising metrics by academic year was; two data sets were subsequently created. For alternative providers, imputation, standardisation and benchmarking were not considered appropriate; however, four data sets were created in order to examine whether predicting the outcome of specific review questions, rather than the overall outcome, would prove beneficial.

With the data sets complete, the next stage was to perform the statistical analysis and determine which metrics, if any, could have predicted the outcome of QAA reviews. The statistical methods selected for this task are discussed below.

4.3. Statistical Methods

To identify a subset of metrics that best predicts a categorical outcome one would typically use logistic regression and a variable selection/reduction method. However, this approach will not work when, as is the case for this analysis, the number of cases (reviews) is less than the number of predictors (metrics) (Kuhn and Johnson, 2013; James *et al.*, 2013). Instead, modern machine-learning techniques are required. This section begins with an introduction to regression and variable selection methods and the issues that prevent them from being effective for this analysis. The requirements for the predictive models and the details of the machine-learning method which best meets them – the *elastic net* approach - are then discussed. Finally, the methods used to evaluate the models developed using the *elastic net* approach are discussed.

4.3.1. Classical Statistical Modelling Techniques

This study aims to select a subset of variables to predict a categorical outcome. Moreover, it aims to do so using a data-driven approach, as discussed in chapter three, to provide the best possible model. The standard approach to achieve this aim would be to use logistic regression and some form of variable selection method. However, this approach was designed for low-dimensional data-sets where the number of observations is far greater than the number of possible independent variables (predictors). That was not the case for this study. With the exception of the alternative provider data sets the number of QAA reviews in each data set was less than the number of metrics available. For the alternative provider data sets, the volume of metrics still made a machine-learning approach appropriate however. Whilst the standard logistic regression approach therefore cannot be used, it is necessary to understand the theory behind this approach, and the problems it is caused by high-dimensional data sets such as the ones we have for this study, to inform the discussion on machine-learning techniques.

Ordinary least squares regression is designed to predict continuous variables such as a person's weight. The standard *OLS* regression model:

$$P(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K$$

where $P(Y|X)$ is the predicted value of the outcome *Y* given the value of the *i* constants β and predictors x is not suitable for this analysis for two reasons. *First*, the outcome is categorical and therefore the requirement of homoscedasticity (that the error terms arising from a variable are normally distributed) is breached. *Second*, the outcome is non-numeric and even if the categorical variables were coded with a number the output would be flawed as false information would have been ascribed to the data (Field *et al.*, 2012).

Logistic regression solves these two problems. Instead of predicting a single value, *Y*, logistic regression modifies the ordinary least squares approach and predicts the *probability* of each possible outcome occurring given specific values for the independent variables. The output can then be regarded as:

$$P(Y|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K)}} = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K)}}$$

(Liao, 1994).

To be able to build a full, multivariate logistic regression model to predict the outcome of QAA reviews it must be determined which metrics to include in the model. To do so variable selection methods are used. There are two main options available. Stepwise selection (Akaike, 1974) has proved popular historically but its use is increasingly being discouraged. Indeed, Harrell (2001) suggests that had it been developed today, the stepwise selection approach would not have been accepted due to its poor performance. Moreover, its use is made practically impossible by the volume of potential predictors. Considering only the smallest data set, alternative providers, there are still 41 metrics meaning there are a possible 2^{41} = 2,199,023,000,000 potential first-order models alone. Hosmer Jr *et al.* (2013) recommend the *purposeful selection* approach which starts with the univariate analysis of every metric and retains those with a Wald test statistic with a p-value of $p<0.25$ (this is higher than the accepted significance threshold of 0.05, because studies by Bendel and Afifi (1977) and Mickey and Greenland (1989) show that selection based on the more traditional threshold can lead to the exclusion of variables known to be important) and continues with careful step-by-step model development on a predictor-by-predictor basis. Whilst a valid approach, it is also clearly infeasible for scenarios with such a large number of predictors.

Aside from both these methods being impractical and often inaccurate, they both have another failing which is that they will *overfit* the data in scenarios such as this where the number of independent variables (metrics) is greater than the number of cases (reviews) (Kuhn and Johnson, 2013). *Overfitting* occurs when the model developed doesn't just account for the general patterns in the data but it also learns all the unique random variation associated with each case and builds this into its predictions. The result is that the model predicts perfectly the outcome of the cases that were used to develop it, but produces grossly inaccurate predictions for any new data. If one were to predict the party allegiance of British Prime Ministers based on their personal characteristics, an overfit model may identify that Britain's two female Prime Ministers to date were Conservative and therefore predict with 100% certainty all future female Prime Ministers will be Conservatives. Clearly this is nonsensical, the model has simply, and incorrectly, learned a

specific fact based on minimal data and applied it to all future cases resulting in inaccurate predictions. With enough metrics, often surprisingly few, it is easy to describe a data set perfectly and this is overfitting.

To illustrate the risk of overfitting, a model containing just the 12 metrics in Table 4.6 below is able to predict with perfect accuracy which of the 184 HEIs in the imputed data set would be 'unsatisfactory' and which would be 'satisfactory'. Moreover, as shown in Figure 4.5, it predicted that those that were 'unsatisfactory' had a 100% probability of being so, and those that were 'satisfactory' had a 0% chance of being 'unsatisfactory'; despite it being impossible to deduce any causal relationship between most of the 12 metrics and quality assurance matters.

Metric code	Metric description
STU018_Ca2	The two-year change in the number of full-person equivalent HE students who qualify with lower second class honours
APL006_Ca2	The two-year change in the proportion of successful applicants whose age is known who are aged 25 & above
STA054_Abs	The proportion of staff (FTE) whose nationality is known who are of "Other-EU" nationality
STA062_Ca1	The one-year change in the proportion of staff (FTE) who are principally financed by the institution.
UCA012_Ca1	The one-year change in the institutional distribution of accepted applicants by domicile Other EU
APL004_Cp2	The two-year percentage change in the proportion of successful applicants whose age is known who are aged 20 & under
UCA023_Ca1	The one-year change in the market share of applications by age 25 years & above
STU077_Ca1	The one-year change in the absolute number of HE students (FPE) who are Other European Union domiciles
APL015_Ca1	The one-year change in the proportion of successful applicants whose domicile is known who are non-EU domiciles
KFI020_Abs	The percentage ratio of contribution from research grants & contracts to research grants & contracts income
DLH012_Abs	The proportion of UK domiciled total leavers who obtained qualifications through full-time study and were reported as unemployed 6 months later
RES020_Ca1	The one-year change in the funding council recurrent grant for research (£000s)

Table 4.6: The 12 metrics that comprise the exemplar model which can describe the HEI data perfectly but make poor predictions with new data.

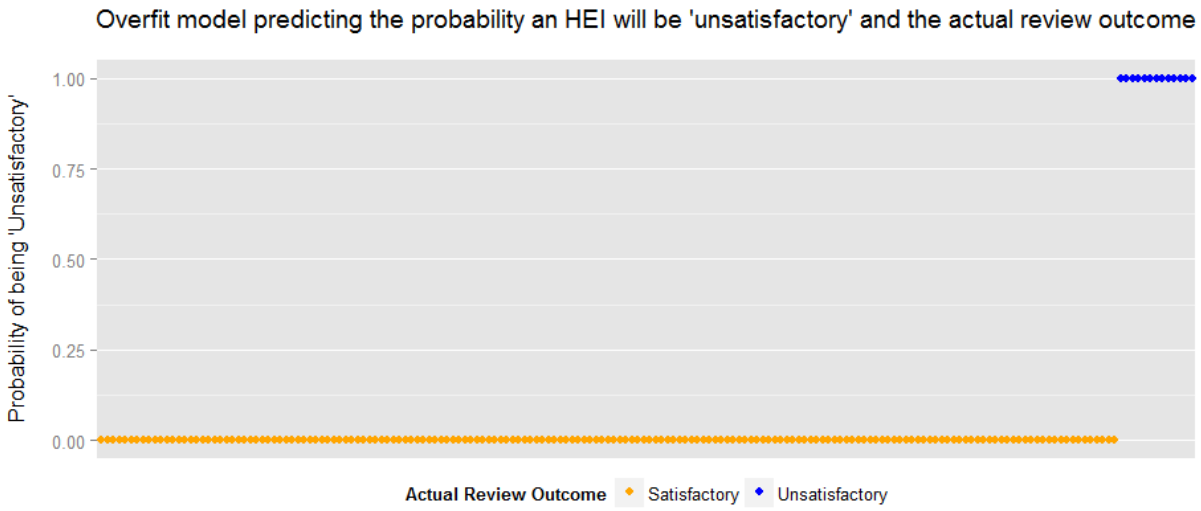

Figure 4.5: The actual outcome and predicted probabilities of those outcomes for the 184 reviews in the HEI data set using the overfit model.

Whilst this model describes the data perfectly, applying it to the 2012/13 data results in predictions of absolute certainty that 18 HEIs would receive an 'unsatisfactory' judgement following the publication of that data. This is more than have been deemed 'unsatisfactory' in total since 2007. All the remaining HEIs have a predicted 0% chance of being deemed 'unsatisfactory'.

The results do not indicate that the QAA could have done a perfect job of risk assessment if armed with this model. On the contrary, the reason this model has been able to so accurately describe the data with which it has been developed, getting each prediction perfectly correct for actual reviews, yet seems to make no logical sense when applied to plausible hypothetical data is because it has 'overfit' the data. The near perfect explanation of the variance is because not only has the model picked up any potential predictive signal amongst the data set but it has also perfectly described the statistical noise surrounding this signal.

In part to address the problem of overfitting, a number of new methods have been developed in the past decade to make it possible to develop effective models when there are many more metrics than observations. Before the requirements of any model developed to predict QAA review outcomes as part of a risk-based approach are considered and the machine-learning approach that best meets those requirements is selected, this study does make use of classical logistic regression in one way which is discussed below.

4.3.2. Univariate Analysis

Before we develop any model, we can run a preliminary analysis of the data to gain an understanding of the relationship between individual metrics and the outcome of the QAA

reviews. This is achieved by running a simple logistic regression between each individual metric and the outcome of the QAA reviews. This does not inform the final, multi-metric model, but is simply designed to give an overview of the data set including which individual metrics appear to have a strong relationship with the outcome of QAA reviews. There are no concerns about overfitting the data or selecting the correct variables when each analysis concerns only one metric.

Part of the output of each of these individual logistic regression models is a *p-value* for the metric used in that regression. A *p-value* is "the probability, if the null hypothesis were correct, of getting as extreme, or more extreme, a result" (Lander, 2014). In the case of logistic regression, the null hypothesis being tested is that the regression coefficient for the parameter in question is zero – i.e. that metric has no bearing on the outcome. So when we obtain a *p-value* of 0.05 this is the equivalent of saying that the probability of the metric in question having no relationship with the outcome is just 5%, or one in twenty. Likewise, the probability that a metric with a *p-value* of 0.25 having no relationship with the outcome is 25%, or one in four. Select metrics, or group of metrics, with a *p-value* of less than the standard 0.05 threshold and the looser 0.25 threshold suggested by Hosmer Jr *et al.* (2013) are discussed prior to each model being developed for each of the three provider types.

It is worth noting at this stage that there are a number of intricacies associated with performing regressions using only one predictor when others may also be relevant. Spurious relationships may arise due to the model failing to account for a *confounding* factor, for example a highly correlated predictor, which is not accounted for in the multivariate model. James *et al.* (2013, p.136) illustrate the phenomenon with credit card default rate data:

> "Students tend to hold higher levels of debt, which is in turn associated with higher probability of default. In other words, students are more likely to have large credit card balances, which … tend to be associated with high default rates. Thus, even though an individual student with a given credit card balance will tend to have a lower probability of default than a non-student with the same credit card balance, the fact that students on the whole tend to have higher credit card balances means that overall, students tend to default at a higher rate than non-students. This is an important distinction for a credit card company that is trying to determine to whom they should offer credit. A student is riskier than a non-student if no information about the student's credit card balance is available. However, that student is less risky than a non-student with the same credit card balance!"

Any counterintuitive relationships could be very revealing, the result of pure chance, or the result of confounding whereby review outcomes in the metric in question are not directly related but

linked by a common variable in the same way that there is no direct link between sunglasses and ice cream but sales of both will be correlated as they are both linked to sunshine. Each relationship will therefore be discussed to determine why it may exist.

4.3.3. Model Requirements and Machine-Learning Approaches

Model development is not a mechanistic process. There is rarely a single 'best' model which can be developed or selected and the total reliance on quantitative measures of fit is rightly discouraged. When developing any model there is a need to balance predictive or descriptive power with simplicity, interpretability and ease of use. In the case of data-driven, risk-based quality assessment of higher education providers, any model developed needs to be of practical use to the QAA or its successor agency. This narrows down the possible statistical approaches that can be used.

First, some situations may require the model to simply describe a binary outcome, did a visitor to a website buy a recommended product or not? Other situations require the more complex prediction of the *probability* of an outcome, what is the likelihood a review of a provider will result in an 'unsatisfactory' judgement? When using the output of a model to prioritise reviews of higher education providers there is a clear advantage in knowing not just whether each provider is forecast to be 'satisfactory' or 'unsatisfactory', but the probability of each outcome occurring. Each provider can then be prioritised in line with how likely it is that they will be judged 'unsatisfactory'.

Second, some models, such as those designed to forecast share prices, can automatically access thousands of data points without any need for human interaction. Other models however, such as those working with multiple, non-public data sources requiring manual data processing will be limited in the volume of data they can use by the resources available to source, process and load that data. For higher education providers the data is not readily available in a way that can be automated. Instead, any model would be reliant on humans to maintain it and it is not feasible, especially as part of a cost effective, burden reducing approach, to have them load thousands of metrics on a regular basis. Moreover, it is unlikely that these thousands of metrics will all prove useful. The goal is therefore to derive a "basket of data" (BIS, 2011, 3.19) that can be monitored. An approach that performs variable selection, that is it eliminates the specific metrics which when combined have no predictive value, while retaining the specific metrics that do, is required.

Third, in some scenarios there is little need to explain how a prediction was reached but a great need for the model to be as accurate as possible. It is not necessary for a weather forecaster to be able to explain why the thousands of factors and measurements feeding into their model

predicts a given chance of rain on a scheduled launch day for a NASA mission, it is however important given what is at stake that the predictions are as accurate as can be. Other scenarios however do require the models to be accessible, even if this is at the expense of accuracy. A higher education provider being reviewed by the QAA whilst similar institutions are not prioritised is unlikely to accept being singled out without an explanation as to why (Schutt and O'Neil, 2013; James *et al.*, 2013). Moreover, reviewers are more likely to be confident that they are targeting their efforts appropriately if they can understand why they have been sent to review a specific provider.

These three criteria rule out complex 'black box' methods such as *support vector machines* which can deliver more accurate estimates than others, but do so with highly-complex data transformations (Hastie *et al.*, 2011; James *et al.*, 2013). Similarly, *boosted tree* methods which run multiple different models and aggregate them to produce a final prediction not easily reproduced or interpreted by humans will not be suitable (Raschka, 2015; Coelho and Richert, 2015). The criteria also rule out *K-nearest neighbour* models which are very straightforward: they simply look at historic data and find the *K* providers whose metric performance is closest to the provider of interest and take a weighted average of the outcome of their reviews. This approach, however, only produces a classification outcome, a single prediction of whether the provider would be 'satisfactory' or 'unsatisfactory' but will not accompany this with a probability (Gutierrez, 2015; Kuhn and Johnson, 2013; Raschka, 2015).

The statistical approach which best fulfils the requirements of a data-driven, risk-based approach to quality assurance in higher education is the *elastic net* approach. The *Elastic Net* is a recent development which produces a penalised logistic regression model by merging two components: *ridge* regression to stabilise the model coefficients and protect against highly correlated predictors (Park and Hastie, 2008; Eilers *et al.*, 2001) and *lasso* regression for selecting predictors for inclusion in the model.

Logistic regression produces the model which minimises the log-likelihood, that is the difference between predicted outcomes and the actual outcomes in the data set. *Ridge* regression regularises this model by seeking to minimise the sum of the log-likelihood plus a penalty factor comprising the sum of the squared predictor coefficients:

$$\log L(p) + \lambda \sum_{j=1}^{p} \beta_j^2$$

Any extreme values are therefore only permitted if they result in a significant reduction in the log-likelihood, i.e. they significantly improve the accuracy of the model. Whilst this shrinks the predictors, none are set to zero, i.e. no predictors are eliminated, unless $\lambda \to \infty$ which makes using and interpreting the model near impossible. *Lasso* regression overcomes this issue by minimising:

$$\log L(p) + \lambda \sum_{j=1}^{p} |\beta_j|$$

This is similar to *ridge* regression but, rather than utilising an ℓ_2 penalty, utilises an ℓ_1 penalty which forces some of the coefficient estimates to equal zero and thus their corresponding predictors are not included in the model (James *et al.*, 2013; Kuhn and Johnson, 2013). The dynamic blending of these two approaches results in the *elastic net* which operates by minimising:

$$\log L(p) + \lambda \left[(1-\alpha) \frac{1}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right]$$

Where λ is the parameter controlling the shrinkage (variable selection) and α is the parameter controlling the proportion of *ridge* versus *lasso* regression used ($\alpha=0$ will result in complete *ridge* regression and all parameters being included, whilst $\alpha=1$ will result in complete *lasso* regression) (Lander, 2014). The *Elastic Net* approach in effect performs the same role as the 'best subset selection' method; however, it does so in a computationally feasible way given large numbers of variables.

The *elastic net* approach also benefits from fitting models using *k-fold cross validation*. This is where the observations of the dependent variable, in this case the QAA reviews, are divided into *k* similarly-sized groups (or *folds*):

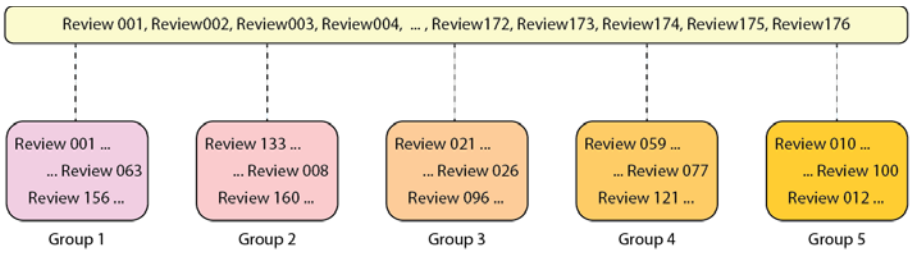

Figure 4.6: An example of dividing reviews into *k*=5 groups for k-fold cross validation.

Each possible combination of *k-1* groups is then used to develop the model and the *log-likelihood*, a measure of the difference between the actual outcomes and those predicted by the model, is calculated when applied to the remaining group.

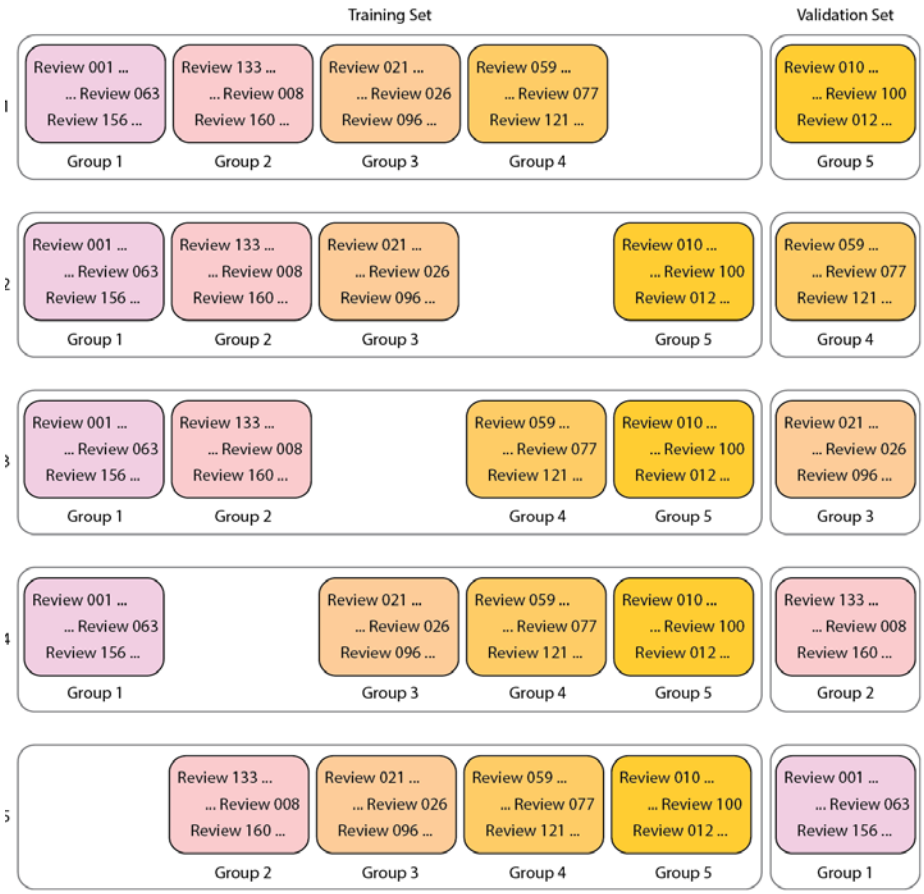

Figure 4.7: Each combination of the 5 groups used in for 5-fold cross validation model development.

This results in *k* cross-validation error scores, one from each training set comprising *k-1* groups, which are then averaged to calculate the overall cross-validation error. The best model can then be selected from those with the smallest cross-validation error. This approach significantly reduces overfitting: at no point is all the data contained in one iteration of model development so all the 'statistical noise' cannot be learned, and any predictions which learn the 'statistical noise' of the group containing *k-1* folds will perform very poorly when validated against the remaining fold and therefore will not be suggested as a valid model.

In summary, the 'dimensionally-cursed' data means that a *machine-learning* approach was required for this study. There are a number of such approaches that could have theoretically been adopted; however, many of these could never work in practice as part of a risk-based approach to quality assurance in higher education. Having considered the criteria a machine-learning approach would be required to fulfil, the *elastic net* approach utilising *k-fold cross validation* was selected. The next stage is to examine how we will evaluate the models produced by this approach.

4.3.4. Model Evaluation

This study makes explicit provision for evaluating the models described in two ways. *First*, how the developed models perform when describing the data that was used in their development is evaluated. *Second*, how well the model performs when predicting the outcome of reviews based on new data is evaluated. The evaluation approach adopted makes use of Receiver Operating Characteristic (ROC) curves, the application of the models to alternative data sets, and, for alternative providers, holding some data back from model development specifically for testing.

4.3.4.1. Testing How Well the Model Describes the Data With Which it Was Developed

For validation purposes, it is useful to undertake a summary analysis of model predictions in terms of true and false positives and negatives: i.e. to examine the trade-off between successfully predicting 'unsatisfactory' reviews (true positives) and predicting 'unsatisfactory' reviews when the review was in fact 'satisfactory' (false positives). This is achieved with two complementary approaches. *First*, a *Receiver Operating Characteristic* (ROC) curve is used.

		Actual Review Outcome	
		Satisfactory	Unsatisfactory
Predicted Review Outcome	Satisfactory	True Negative	False Negative
	Unsatisfactory	False Positive	True Positive

Table 4.7: An illustration of true and false positives and negatives. When predicting unsatisfactory reviews, the correct prediction of an 'unsatisfactory' review is regarded as a 'true positive'. The incorrect prediction of a 'positive' result, i.e. the HEI will be judged 'unsatisfactory' is deemed a 'false positive'.

Different scenarios have a different acceptable balance between the true and false positives. The QAA may not be willing, or have the resource, to review 50 HEIs which turn out to be 'satisfactory' for the sake of reviewing one that is 'unsatisfactory'. Conversely, the security services may well be willing to actively monitor 50 individuals who subsequently turn out to be innocent for the sake of stopping one genuine threat. The ROC curve demonstrates the impact on true-positive and false-positive rates of triggering reviews based on different predicted probabilities of a review resulting in an unsatisfactory judgement.

The ROC approach allows us to go beyond the simple classification approach, i.e. did it predict a probability of failure greater than 50% or not, and adjust our thresholds which can result in a more useful and effective model (Fawcett and Provost, 1997; Provost *et al.*, 1998). Moreover, the ROC approach is especially useful in cases such as this where there are skewed class distributions, e.g.

there were a far greater number of ‘satisfactory’ reviews than ‘unsatisfactory’ reviews (Fawcett, 2006). The ROC also has the benefit of being intuitively graphically displayed.

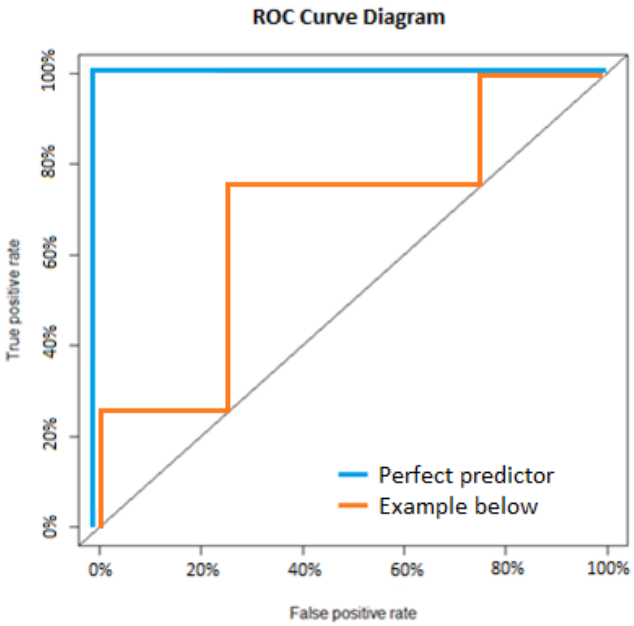

Figure 4.8: An exemplar ROC curve.

Each point on an ROC curve indicates the rate of ‘true positives’ at varying thresholds for the acceptability of ‘false positives’, in the case of this study the varying probabilities of an ‘unsatisfactory’ review which would prompt a QAA review. This is best explained by an example. Consider a model which predicts the probability that 8 HEIs will be judged ‘unsatisfactory’ by the QAA were they to be reviewed:

Review priority	Predicted probability of ‘Unsatisfactory’ judgement	Actual outcome	True positives (TP)	True negatives (TN)	False positives (FP)	False negatives (FN)	True positive rate = TP/(TP+FN)	False positive rate= FP/(FP+TN)
1	80%	Unsatisfactory	1	4	0	3	0.25	0
2	70%	Satisfactory	1	3	1	3	0.25	0.25
3	65%	Unsatisfactory	2	3	1	2	0.5	0.25
4	58%	Unsatisfactory	3	3	1	1	0.75	0.25
5	52%	Satisfactory	3	2	2	1	0.75	0.5
6	50%	Satisfactory	3	1	3	1	0.75	0.75
7	45%	Unsatisfactory	4	1	3	0	1	0.75
8	37%	Satisfactory	4	0	4	0	1	1

Table 4.8: Exemplar model predictions and the outcome of using different threshold probabilities to trigger a review.

The logical approach would be to order the HEIs by their predicted probabilities of an ‘unsatisfactory’ review from largest to smallest. If the QAA were to take a predicted probability of

being judged ‘unsatisfactory’ of 80% as the threshold to trigger a review only one would take place. That HEI would be judged ‘unsatisfactory’ so the decision to review it would be correct meaning the model predicted one ‘true positive’. Of the seven HEIs not reviewed, four would have been judged ‘satisfactory’ and therefore not reviewing them was the correct decision meaning the model produced four ‘true negatives’. There would however have been three ‘unsatisfactory’ HEIs not reviewed meaning there were three ‘false negatives’. As no ‘satisfactory’ HEIs would have been reviewed there would be ‘false negatives’. This would result in a true positive rate – the proportion of ‘unsatisfactory’ HEIs reviewed - of one out of four, or 0.25, and a false positive rate – the proportion of ‘satisfactory’ HEIs reviewed – of none in four, or 0.

Now consider what would happen if the QAA were to lower the threshold predicted probability required to trigger a review to 70% meaning two HEIs would be reviewed. Now one ‘unsatisfactory’ HEI would be correctly reviewed (one true positive) and one ‘satisfactory’ HEI would be incorrectly reviewed (one false positive). Furthermore, three ‘unsatisfactory’ HEIs would incorrectly not be reviewed (three false negatives) and three ‘satisfactory’ HEIs would correctly not be reviewed (three true negatives). This would result in a true positive rate of 0.25 and a false positive rate of 0.25.

Each lowering of the threshold predicted probability required to trigger a QAA review alters the true and false positive rates and is represented by a point on the ROC curve. If the additional review triggered by the lowering of the threshold is of an ‘unsatisfactory’ provider, the true positive rate will increase and the false positive rate will remain the same, resulting in the line moving upwards. Conversely, if the additional review triggered by the lowering of the threshold is of a ‘satisfactory’ provider, the true positive rate will remain the same and the false positive rate will increase, resulting in the line moving right. The perfect predictive model would therefore have a curve that was a vertical line hugging the y-axis – indicating that all the reviews with the highest predicted probability of being ‘unsatisfactory’ did indeed result in an ‘‘unsatisfactory’ judgement - and then, once all the true positives have been correctly predicted, a horizontal line from the top of the y-axis – representing all the reviews below a threshold predicted probability of being judged ‘unsatisfactory’ were not judged ‘unsatisfactory’.

The diagonal line represents the worst-case scenario: it is the path the results of a predictive model would follow if it had no predictive ability whatsoever and could not distinguish between true and false positives. It is of course possible a model could be wrong more than 50% of the time, and therefore its performance would be shown by a line beneath the diagonal; however, in such cases one merely needs to do the exact opposite of what the model suggests to achieve predictions

better than chance. The closer the resemblance of the ROC curve to the perfect fit line, the better the model is at describing the data (Verostek, 2014). How close this fit is, and thus how well the model describes the data, can be measured by the area under the curve (AUC) with values close to 0.5 indicating no discriminatory power and values close to 1 indicating an excellent fit. Using the AUC to compare models however should be done with caution as two ROC curves may well have the same AUC value but cross several times.

The *second* complementary approach used to evaluate how well the model describes the data used in its development is a visual inspection of the model's predictions and the actual review outcomes. This is achieved with a table similar to Table 4.8 above and a graphical representation akin to Figure 4.9 below.

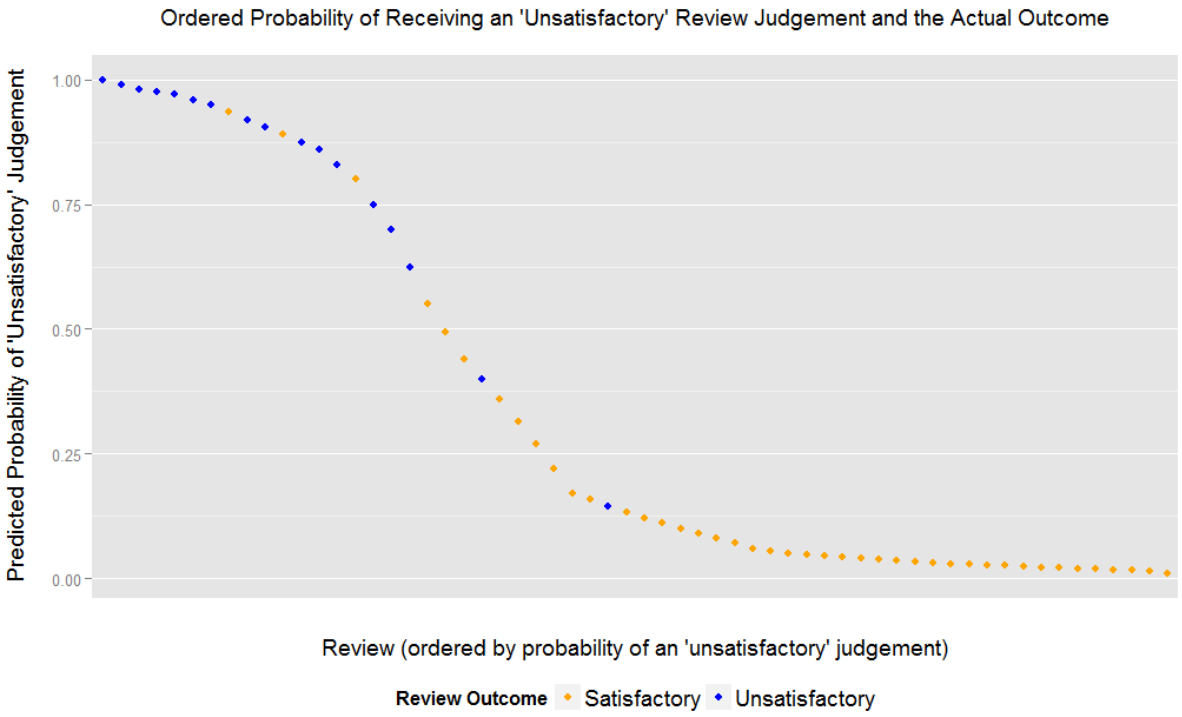

Figure 4.9: An example of the predicted probabilities from a successful model and actual review outcomes with the reviews ordered by the predicted probability of being 'unsatisfactory'.

Figure 4.9 provides an example of a distribution that would result from a successful model: there is a clear range of predicted probabilities with the majority of reviews predicted to be either 'unsatisfactory' or 'satisfactory' with a high degree of certainty. By definition, we would expect nine out of 10 reviews predicted to be 'unsatisfactory' with a probability of 90% to result in an 'unsatisfactory' judgement and, by extension, would expect to see one out of 10 result in a 'satisfactory' judgement. It is to be expected therefore that, even with a very accurate predictive model, a proportionate number of 'satisfactory' reviews will be found to the left-hand side of the

plot amongst the 'unsatisfactory' reviews, and likewise a proportionate number of 'unsatisfactory' reviews will be found to the right-hand side of the plot amongst the 'satisfactory' reviews.

4.3.4.2. Testing How Well the Model Performs Predicting Additional Cases

When developing a predictive model an analyst must decide how to 'spend' their data, i.e. how much of it should be dedicated to developing (or 'training') the model and how much of it should be held back, not used in the model development, and instead used to test the predictive performance of the developed model. When the number of cases, or 'reviews' in this study, is large setting aside a portion of the data allows for an unbiased evaluation of the final model. The trade-off however is that not all of the known information is used to develop the model in the first place.

A benefit of the cross-validation approach used in this study is that withholding testing data for the purposes of assessing model accuracy is not necessary. Indeed, multiple studies have shown the limitation of approaches based on withholding data for testing (see for example (Martin and Hirschberg, 1996; Molinaro *et al.*, 2005)). Hawkins *et al.* (2003) conclude that such approaches "do not match the cross-validation itself for reliability in assessing model fit." Where holding back testing data is still useful however is as a sense check – even if we have a well-performing, cross-validated model it can be reassuring to know whether it is an effective predictor in the real world by seeing its application to new data.

Due to the different nature of the three provider types, the timing of the analyses and the availability of data featured in the developed models, a tailored approach was adopted to provide this sense check for each. These are discussed in turn below.

4.3.4.2.1. HEIs

For HEIs the number of reviews is small and the proportion of 'unsatisfactory' judgements the study aims to predict is low. A strong case can therefore be made against splitting the data as each case contributes significantly to the model-building process. Instead, another way to sense check the model is required.

One option that had to be ruled out was waiting for further HEI reviews to be conducted and comparing the outcome to what the model predicted. HEIs are only reviewed once every six years and, with nearly all HEIs reviewed in the first four years of the current six-year cycle, there was no prospect of a robust sample of reviews building up for sense checking the model. There are however two approaches available. *First*, although certainly not robust or definitive, as a sense check we can apply HEI models to the data sets concerning the 2012/13 academic year and can

gain an approximation of which HEIs are predicted as most likely to be 'unsatisfactory' now. The results can then be assessed to see whether they appear highly or moderately plausible, or totally implausible, using other publicly-available (non-QAA) judgements on the sector. Data confidentiality prevents any HEIs being named as part of this analysis, but it was achieved by grouping institutions using their (somewhat contested) 2014 Guardian University Rankings (The Guardian, 2013). *Second,* by applying the latest data available at a specific point in the time in the past we can see which HEIs would have been at the top of the QAA's priority list and where on that priority list those providers that were actually reviewed around that time were positioned. The results of both these checks must be regarded with caution but provide two basic sense checks on the likely validity of the model(s).

4.3.4.2.2. FECs

For FECs, as with HEIs, the number of reviews is small and the proportion of 'unsatisfactory' judgements the study aims to predict is low. Again, a strong case can therefore be made against splitting the data using an alternative method to sense check the model.

Two approaches were adopted. *First*, over 50 reviews have taken place since the original review data was gathered and the updated metrics which comprise the predictive models are available. This means the new predictions created by the models can be directly tested against the outcome of the new reviews using the ROC and visualisation methods described above in 4.4.4.1 above. *Second,* as with HEIs we can apply the model to existing performance data at a specific, historic point in time to provide a view of what the QAA would have been presented with at the time had they been using the model, rather than the perfect hindsight we have now, and seeing where on the distribution of predictions those FECs reviewed within one year were placed.

Unlike HEIs, FEC rankings are not available nor do most people know the reputation of each FEC's HE provision; however, this is of little concern as testing on actual reviews is far superior to using reputational rankings as a proxy.

4.3.4.2.3. Alternative Providers

As with FECs, new reviews of alternative providers have taken place since the data was gathered for this analysis. These reviews could not be used to test the model however as the data necessary to predict the outcome of these new reviews – the providers' financial accounts submitted to Companies House - are prohibitively resource intensive to obtain. The two other testing approaches adopted for HEIs and FECs, assessing predictions against provider rankings and

applying the model to a fixed historical point, would not be possible as no such rankings exist and many young providers do not have historic financial information respectively.

For the alternative provider analysis, however, there is a greater number of reviews and 'unsatisfactory' judgements than for HEIs or FECs and we can set aside some data for model testing without a major impact on the model development process. Prior to the development of each alternative provider model, a representative sample of reviews was selected using the caret package in R (Kuhn, 2008) and set aside for testing while the remaining reviews were used to develop the model. The model was subsequently tested using this withheld data and the ROC and visualisation methods described above.

4.3.4.3. Defining 'Success'

Few would disagree that a model which performs as shown in Figure 4.9 is a success. Were the QAA able to conduct reviews in order of their likelihood to result in an 'unsatisfactory' judgement and halt their review activity at the optimal point following the final 'unsatisfactory' review, all 17 'unsatisfactory' reviews could have been conducted with only 12 'satisfactory' providers being reviewed. This would be an error rate – i.e. the proportion of reviews conducted that turned out to be 'satisfactory' – of 12/(12+17) = 41%. Providers prioritised for review were almost 50% more likely to be 'unsatisfactory' than not and 31 'satisfactory' providers would have been spared the burden of a review.

As performance of the model diminishes, however, perceptions of its success become more subjective. Discussion of model success must consider three factors. *First*, how well does the model differentiate the 'unsatisfactory' providers from the 'satisfactory' providers? This will be defined by the shape of the curve: a sweeping 'backwards S' shape curve covering a range of probabilities will differentiate well, while a flatter curve, or one over a narrower range of probabilities, will perform worse. *Second*, and, not withstanding outliers, connected to the shape of the curve is how many 'satisfactory' providers could have been spared their review had QAA stopped at the optimal point? As noted above, there is no objective definition of an acceptable curve. It will depend on the tolerance of QAA and other key stakeholders for false negatives and the number of reviews QAA are resourced to conduct in a given time period. Both the models' ability to differentiate providers and the shape of the curve will affect the model's error rate. *Third*, how well does the model perform when applied to new data? A model which describes the data with which it was developed very well, and with a low error rate, will be of little use if it does not work when applied to new data.

In summary, how well each model describes the data with which it was developed was assessed using an ROC curve and the associated AUC value, and visualisation of each model's output. This approach assessed how many 'satisfactory' and 'unsatisfactory' providers would be reviewed as the predicted probability of a provider being judged 'unsatisfactory' required to trigger a review was lowered. Each model was also tested to see how well it performs on new data although the specifics of this vary by sector due to the nature and availability of their data. For HEIs the models' predictions were compared with (somewhat contested) rankings and their performance at a specific point in time is evaluated. For FECs more recent reviews and metric data became available since the models were developed which was used to provide a real-life evaluation of how the model would perform. For alternative providers a representative subset of the data was held back from the model development process and used to test the final model. It is the combination of these tests which define the success of a model; however, that success is often subjective.

4.3.5. Summary

The classical statistical methods for modelling categorical outcomes were not suitable for this study. Having a greater number of metrics than reviews mean that using such approaches would result in a model which is greatly overfit; i.e. a model that doesn't just account for the general patterns in the data but it also learns all the unique random variation associated with each case and does a poor job at predicting new cases as a result. Instead, a machine-learning approach was required. Having considered the criteria a machine-learning approach would be required to fulfil, the *elastic net* approach utilising k-fold cross validation was selected. Each model developed using the elastic net approach was evaluated to determine how well it described the data with which it was developed and how well it predicted outcomes based on new data.

4.4. Conclusion

The aim of this chapter has been to detail the methods used to determine whether metrics could have predicted the outcome of QAA reviews and therefore have been used to prioritise reviews according to risk.

First, I have detailed the dependent and independent variables used in this study and that, due to the nature of data and the improved model accuracy that will result, separate models will be developed primarily at the overall review outcome level.

Second, I have detailed the steps taken to prepare the data ready for developing the models. This requires calculating change-over-time metric variants, matching reviews with the latest available

metrics, removing anomalous reviews and metrics, assessing and addressing the missing data including, where appropriate, imputing data, standardising and benchmarking data where appropriate, and removing redundant non-variant and highly-correlated metrics.

Third, I have detailed the statistical modelling approach used, the reasons for its selection, and how the models developed will be evaluated.

The next stage is to detail the application of the methods discussed here to the data sets for the three provider types: HEIs, FECs and alternative providers respectively.

Appendix C – Overfitting

As stated in section 4.4.1 o*verfitting* occurs when a model doesn't just account for the general patterns in the data but it also learns all the unique random variation associated with each case and builds this into its predictions. The result is that the model predicts perfectly the outcome of the cases that were used to develop it, but produces grossly inaccurate predictions for any new data.

An overview of an overfit model was given in section 4.4.1. Below the model is explored further to demonstrate clearly incorrect predictions produced by a model which, on initial fitting, seems to be extremely accurate.

The overfit model contains twelve metrics described below in Table 4.8 and calculates the probability of a QAA review having an 'unsatisfactory' outcome as:

$$P(Unsatisfactory) = \frac{e^G}{1 + e^G}$$

where:

$$G = -397.2 + (1 \times STU018\_Ca2) + (9629 \times APL006\_Ca2) + (-2207 \times STA054\_Abs) + (11630 \times STA035\_Ca1) + (39.3 \times UCA012\_Ca1) + (5436 \times APL004\_Cp2) + (-1059 \times UCA023\_Ca1) + (-1.1 \times STU077\_Ca1) + (5200 \times APL015\_Ca1) + (-0.6 \times KFI020\_Abs) + (14.1 \times DLH012\_Abs) + (0.04 \times RES020\_Ca1)$$

The predictions that result when this model is applied to the HEI data are all flawlessly correct; those reviews that resulted in an 'unsatisfactory' judgement were forecast as being 'unsatisfactory' with absolute certainty (i.e. 100% probability) and those reviews that resulted in a 'satisfactory' judgement were forecast as being 'satisfactory with absolute certainty (i.e. 0% probability of being 'unsatisfactory').

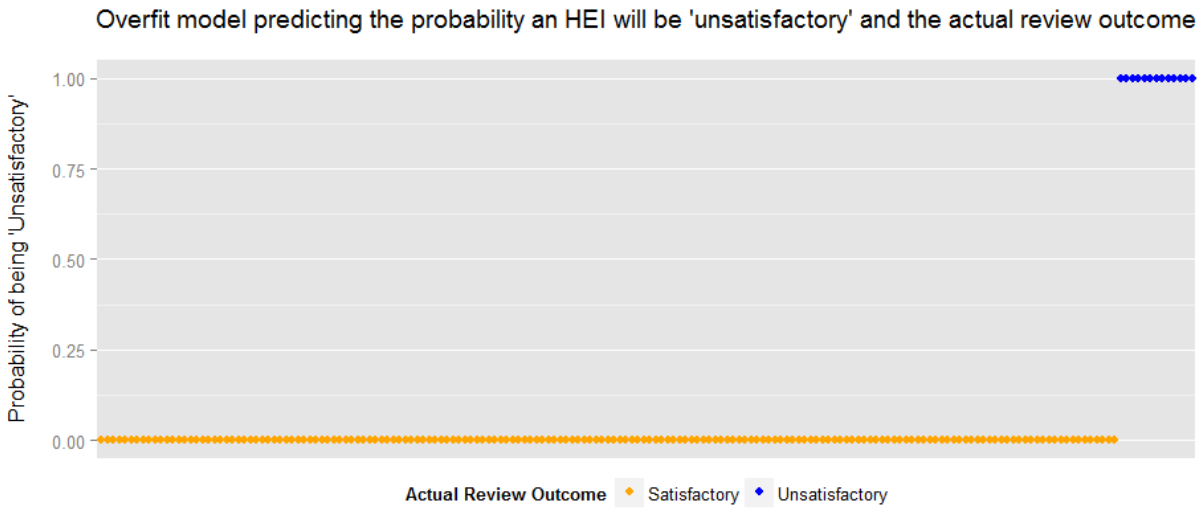

Figure 4.5 (repeated): The actual outcome and predicted probabilities of those outcomes for the 184 reviews in the HEI data set using the overfit model.

Whilst this model describes the data perfectly, applying it to the 2012/13 data results in predictions of absolute certainty that 18 HEIs would receive an 'unsatisfactory' judgement following the publication of that data. This is more than have been deemed 'unsatisfactory' in total since 2007. All the remaining HEIs have a predicted 0% chance of being deemed 'unsatisfactory'. The model perfectly describes the data, including all the errors and variance, but is a poor predictor of future reviews (James *et al.*, 2013). The reason why is best explained by exploring the output of the model for three hypothetical HEIs with plausible and naturally arising values for each of the metrics. These are displayed in Table 4.9:

Metric code	Metric description	HEI A	HEI B	HEI C
STU018_Ca2	The two-year change in the number of full-person equivalent HE student who qualify with lower second class honours	150	0	-300
APL006_Ca2	The two-year change in the proportion of successful applicants whose age is known who are aged 25 & above	0.07	0	-0.03
STA054_Abs	The proportion of staff (FTE) whose nationality is known who are of "Other-EU" nationality	0.42	0.0314	0.01
STA062_Ca1	The one-year change in the proportion of staff (FTE) who are principally financed by the institution.	0.02	0	-0.01
UCA012_Ca1	The one-year change in the institutional distribution of accepted applicants by domicile Other EU	0.3	0	0.15
APL004_Cp2	The two-year percentage change in the proportion of successful applicants whose age is known who are aged 20 & under	0.25	0	-0.1
UCA023_Ca1	The one-year change in the market share of applications by age 25 years & above	0.01	0	-0.02
STU077_Ca1	The one-year absolute change in the number of HE students (FPE) who are Other European Union domiciles	250	0	-100
APL015_Ca1	The one-year change in the proportion of successful applicants whose domicile is known who are non-EU domiciles	0.5	0	-0.25
KFI020_Abs	The percentage ratio of contribution from research grants & contracts to research grants & contracts income	0.25	0.0225	-0.1
DLH012_Abs	The proportion of UK domiciled total leavers who obtained qualifications through full-time study and were reported as unemployed 6 months later	0.02	0.045	0.05
RES020 _Ca1	The one-year change in the funding council recurrent grant for research (£000s)	10000	0	-500

Table 4.9: The hypothetical performance of three HEIs on the 12 measures contained in the exemplar model which describe the data set near perfectly.

HEI A's performance is what many would strive for: amongst other things it has seen an increase in the proportion of successful mature applicants, an increase in the value of research grants from its funding council, and a low unemployment rate for its leavers. Its probability of receiving an unsatisfactory review can be calculated as:

$$\frac{e^{3819.65}}{1 + e^{3819.65}} = 1$$

In other words, the 'desirable' values it obtains on the model variables result in a prediction of absolute certainty – a 100% probability - that the HEI would be 'unsatisfactory' if reviewed.

HEI B has experienced no change whatsoever over the past two years and is performing on the boundary of the lowest quartile for the three metrics evaluating its current performance, rather than its change in performance. Its probability of receiving an unsatisfactory review can be calculated as:

$$\frac{e^{-465.90}}{1 + e^{-465.90}} = 0$$

In other words, by performing relatively poorly but not getting any worse, the HEI is forecast to be 'satisfactory' with absolute certainty.

Amongst other things HEI C has seen a decrease in the number of successful mature applicants, a reduction in the value of their research grants from their funding council, and relatively high unemployment amongst its leavers. Its probability of receiving an unsatisfactory review can be calculated as:

$$\frac{e^{-2852.56}}{1 + e^{-2852.56}} = 0$$

Even though it is performing poorly, indeed *because* it is performing poorly on the select metrics which have overfit the model but do not describe any of the general, meaningful trends that help predict the outcome of QAA reviews, HEI C is still forecast as being 'satisfactory' with absolute certainty.

This model then describes the data with which it was developed very well but performs poorly with the new hypothetical data. Any HEI such as HEI B which does not change very specific aspects of its staff, student and successful applicants mix is predicted as having a 0% chance of being 'unsatisfactory', no matter how poorly they are performing. Any HEI such as HEI A which seems to be improving and performing above average on the three absolute measures in the model is predicted as being certain to be 'unsatisfactory' whereas any HEI, such as HEI C, performing poorly and getting worse is predicted as being certain to be 'satisfactory'. Clearly performing well on certain measures and improving should not be a sign of guaranteed 'unsatisfactory' review outcomes. Likewise, performing poorly on some measures and getting worse should not be a sign of guaranteed 'satisfactory' performance.

This example demonstrates the importance of using a statistical approach which is not susceptible to overfitting and evaluating a model once developed.

Appendix D – Imputation Methods

Imputation, or statistical estimation, can sometimes be used to fill in the missing values in a data set. Imputation can be used where values are not missing for structural reasons - it is not possible to estimate the completion rate of postgraduate taught students at an HEI which does not offer postgraduate taught degrees – but rather where they are 'missing at random' – they should exist but have not been collected or made available for some non-systematic reason. As detailed in section 4.3.8.1 imputation was only deemed suitable for HEIs and for FECs or alternative providers which have much greater uniformity and coverage in their data sets.

There are two possible approaches to imputation. Single imputation uses the information available from one variable, for example it may fill missing values with the mean value for that metric. Multiple imputation utilises additional data and uses this to make an approximation of missing values, for example for if one NSS metric value is missing for an HEI a multiple imputation approach could consider which other HEIs performed most like the HEI with the missing value on the questions for which data is not missing, and use their values for the question that does contain missing data. The *K nearest neighbours* (KNN) method I used for all data sets except one (discussed below), does just that. I used the DMwR package in R (Torgo, 2013) to calculate a weighted average of the five nearest neighbours, defined by their Euclidean distance, for each missing value (Batista and Monard, 2002) . This approach is deemed more robust than single imputation methods as it takes more information into account and produces an HEI-specific estimation of each missing value (Zhao and Cen, 2013).

For one data source, the staffing metrics, it was not possible to use the KNN method due to the number of missing values. The proportion was still low, no more than 15% of the values were missing for any metric; however, this meant an alternative single approach was required. With less than 15% of the values missing the difference between simply imputing the median and the more computationally intensive bootstrapping approach is extremely negligible (Harrell, 2001). The staffing metrics were therefore imputed with the metric median value.

5. Predicting the Outcome of HEI Reviews

The purpose of this chapter is to determine which metrics, if any, could have predicted the outcome of past QAA HEI reviews as part of a robust model and how accurately they could have done so. To do so most effectively and comprehensively four separate questions are explored:

1. Using only complete metrics, could the outcome of QAA HEI reviews have been successfully predicted?
2. With the use of statistical imputation and all comprehensive metrics, could the outcome of QAA HEI reviews have been successfully predicted?
3. With the use of statistical imputation and all in-year standardised, comprehensive metrics, could the outcome of QAA HEI reviews have been successfully predicted?
4. With the use of statistical imputation and all comprehensive, benchmarked metrics, could the outcome of QAA HEI reviews have been successfully predicted?

An overview of the HEI sector and its specific challenges are detailed below. A step-by-step description of the analyses and results is then presented followed by a brief discussion of the findings. The analyses for the first two questions are presented in full detail to ensure readers have a comprehensive understanding of the approach.

5.1. Sector Overview

In 2011/12 there were 164 HEIs in the UK responsible for 2,500,000 HE students. This dwarfed the estimated 674 alternative providers and 253 FECs responsible for 160,000 and 118,000 UK HE students respectively (HESA, 2012; BIS, 2013; HEFCE, 2015a). The HEI sector is unique in being data rich but case poor; there are thousands of HEI-focused metrics available but there have only been 13 incidents of HEIs being judged 'unsatisfactory' by the QAA. Responsible for the majority of HE provision, HEIs are the most compliant of the three provider types by a considerable margin. However, whilst the likelihood of an HEI receiving an 'unsatisfactory' review is low when compared with other provider types, the impact of that failure, either in terms of the number of students affected or the harm done to the reputation of UK higher education, is far greater.

Over the seven-year period covered by the data set (2007-2014) there have been some significant changes to the HEI landscape. Foremost was the near tripling of tuition fees to £9,000 a year from 2012/13 with the aim of encouraging more of a 'market' for higher education. The hoped for competition on fees did not materialise; rather than charging £9,000 only in 'exceptional circumstances' the average full-time fee in 2015/16 is £8,703 (Hope, 2011; OFFA, 2015). Increased fees however resulted in students increasingly seeing themselves as consumers and expecting

more from their universities (Universities UK, 2013). A second major change was the lifting of the student numbers cap. In 2012/13 HEIs were allowed to recruit as many 'AAB' students as they wished, in 2013/14 this was extended to include 'ABB' students, and for the 2015/16 academic year the cap was lifted for all students (Hillman, 2014). Throughout this recent period of change the QAA's approach has remained consistent. Due to the cyclical nature of HEI reviews only a small number were undertaken in 2013 and 2014 and so these any impact is unlikely to have filtered through to the data and have an impact on this study.

5.2. Results – Naturally-Complete Data

Using only naturally-complete metrics, could the outcome of QAA HEI reviews have been successfully predicted?

5.2.1. Initial Data Exploration

The first step was to run a univariate analysis for each of the 753 metrics to gain an understanding of the relationships between individual metrics and the outcome of QAA reviews. Table 5.1 below shows the 14 metrics with a p-value of less than 0.05:

Metric Code	Metric Description	P-value
UFI015_Ca1	The one-year change in the difference between historical cost depreciation & the actual charge for the year calculated on the re-valued amount	0.0058
STU018_Ca2	The two-year change in the number of HE student qualifiers (Full-Person Equivalent - FPE) who are Lower second class honours	0.0123
STA054_Abs	Proportion of staff (Full-Time Equivalent - FTE) whose nationality is known who are of "Other-EU" nationality	0.0159
STA032_Ca1	The one-year change in the number of staff (FTE) who are Principally financed by the institution	0.0160
STU109_Cp1	The one-year percentage change in the number of HE students (FPE) who are "UK LEA mandatory/discretionary awards" funded	0.0179
STU079_Cp1	Proportion of HE students (FPE) who are UK domiciles	0.0214
KFI008_Abs	Percentage ratio of tuition fees & education contracts to total income	0.0221
UFI042_Abs	Balance sheet - Provisions for liabilities and charges - Balance as at 31 July	0.0240
STU085_Cp1	Number of HE students (FPE) who are Non-UK domicile	0.0379
STA031_Ca1	The one-year change in the number of staff (FTE) who are wholly institutionally financed	0.0391
STU077_Ca1	The one-year change in the number of HE students (FPE) who are Other European Union domiciles	0.0421
STU024_Abs	Proportion of all FPE HE Student Qualifiers who achieve a Lower Second Class Honours degree (out of all applicable degrees)	0.0432

KFI020_Abs	Percentage ratio of contribution from research grants & contracts to research grants & contracts income	0.0465
NSS005_Abs	Q5 - The criteria used in marking have been clear in advance	0.0499

Table 5.1: All metrics from the naturally-complete HEI data set with a univariate p-value < 0.05.

With 753 metrics we would expect by chance alone to see 753 x 0.05 = 38 metrics flag at the $p < 0.05$ level and 753 x 0.25 = 188 metrics flag at the $p < 0.25$ level even if there were no meaningful relationships. This is greater than the actual numbers of 14 and 164 metrics respectively that achieved these significance levels; it is far from certain therefore that the data demonstrates any real-world significance rather than just chance relations.

Examining these metrics further suggests that they may not be as much use as hoped. Figure 5.1 below shows the distribution of the metric with the greatest statistical significance, *UFI015_Ca1 - The one-year change in the difference between historical cost depreciation & the actual charge for the year calculated on the re-valued amount*, and whilst four of the 13 reviews which resulted in an 'unsatisfactory' judgement have a metric score of less than 0, the remaining nine do not. Although no HEI that was in the minority of organisations with a metric score greater than 0 was deemed 'unsatisfactory', the majority of 'satisfactory' and 'unsatisfactory' HEIs scored 0. This is not particularly helpful in identifying those that will be deemed 'unsatisfactory'. Moreover, it is not clear why this metric may help identify quality assurance failings.

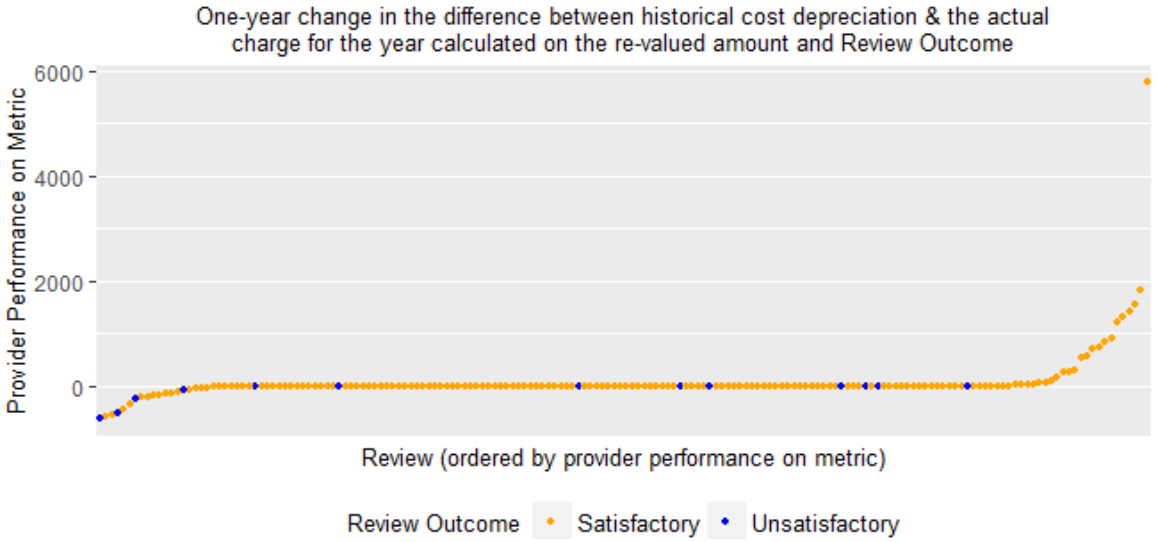

Figure 5.1: A plot of the 'one-year change in the difference between historical cost depreciation & the actual charge for the year calculated on the re-valued amount' prior to each review and the outcome of that review.

Figure 5.2 below concerning the one-year percentage change in the proportion of full-person equivalent (FPE) students who were domiciled in the UK prior to beginning their course exhibits a similar pattern. The HEI with the greatest percentage increase was judged 'unsatisfactory', as were

a number of other HEIs seeing a relatively large percentage increase in the proportion of FPE students who were domiciled in the UK prior to beginning their course. However, these 'unsatisfactory' HEIs are distributed amongst a number of 'satisfactory' HEIs and seven of the 13 'unsatisfactory' HEIs had no change or a slight decrease in the proportion of FPE students who were domiciled in the UK prior to beginning their course. One can imagine concerns over quality, not stemmed by 'unsatisfactory' quality assurance processes, leading to a decrease in international students wishing to study at an institution; however, there are multiple other possible explanations including the opening or closing of courses disproportionately popular or unpopular with international students, or fee changes in specific years making UK HEIs as a whole more or less attractive (such temporal issues would be addressed by the in-year standardisation of metrics used in section 5.4).

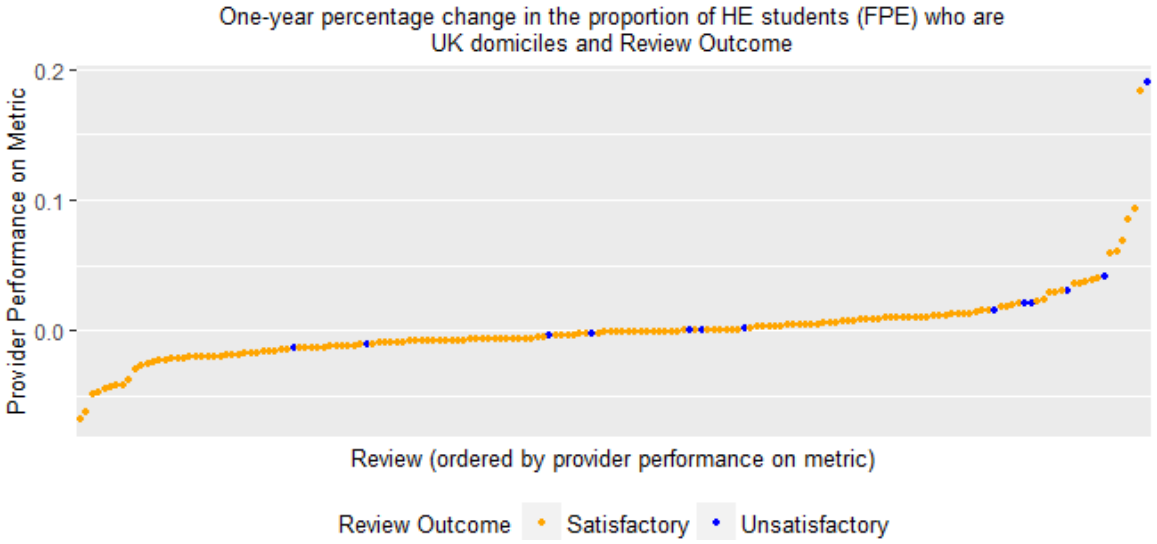

Figure 5.2: A plot of the latest value of 'the one-year percentage change in the proportion of full-person equivalent (FPE) students who were domiciled in the UK prior to beginning their course' prior to each review and the outcome of that review.

The one National Student Survey (NSS) metric to have a p-value of less than 0.05 was *NSS005_Abs - Q5 - The criteria used in marking have been clear in advance* shown in Figure 5.3 below. One can see a connection between marking criteria being made clear in advance and chapter B6 of QAA's *Quality Code for Higher Education* titled "Assessment of Students and the Recognition of Prior Learning" (2012).

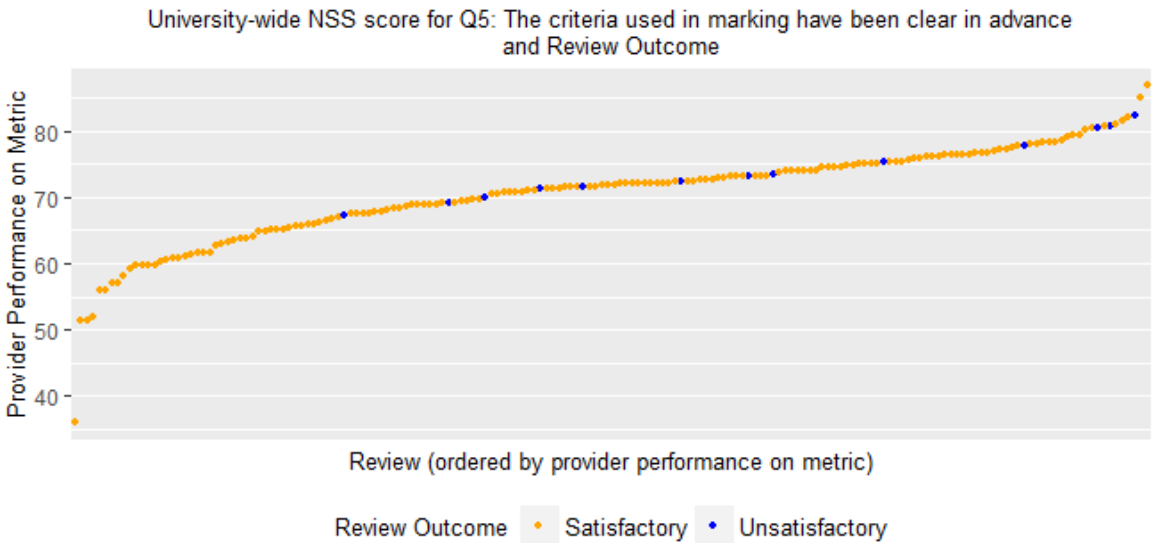

Figure 5.3: A plot of the latest value of 'NSS Q5 - The criteria used in marking have been clear in advance' prior to each review and the outcome of that review.

On closer inspection however the relationship is actually the opposite of what might be expected. Those reviews resulting in an 'unsatisfactory' judgement tend to be preceded by an NSS metric score that is average or better including three in the top 10. The cluster of 'satisfactory' providers on the left of the plot are those that performed worst when students were asked if marking criteria were made clear prior to submission of work.

The inspection of significant individual metrics is not reassuring. There were a substantial number of metrics with significant p-values when modelled in isolation against the review outcomes; however, this represented fewer than we would expect to see by chance and, on closer inspection, many of the results did not seem useful for predictive purposes. The next step is to see if a model can be developed which combines individual metrics, not useful in isolation, into something greater than the sum of its parts.

5.2.2. Fitting the Model

As detailed in the methods chapter the *elastic net* approach requires two tuning parameters: λ to control the shrinkage (variable selection) and α to control the proportion of *ridge* versus *lasso* regression used. These parameters are determined by running the *elastic net* model at various levels of α for which we obtain the minimised cross-validation errors and the optimal values of λ used to achieve these. It is considered best practice to prefer the *lasso* to the *ridge* regression and so only values of $\alpha \geq 0.5$ are considered (Lander, 2014). Both the optimal value of λ (λ_{min}) and the largest value of λ with a cross-validation error that is within one standard error of the minimum (λ_{1se}) are returned. The latter is of interest as the principle of parsimony suggests that we should prefer the simpler model that will be created using λ_{1se} despite it being slightly less accurate. We therefore consider two plots at each stage of model-fitting – the optimum model and the best,

simpler but less accurate model within specified limits – before choosing the preferred model for evaluation.

Figure 5.4 below shows the cross-validation errors obtained for various levels of α when using the λ_{1se} value (top) and the λ_{min} value (bottom). In both cases the cross-validation error is minimised by $\alpha = 0.5$:

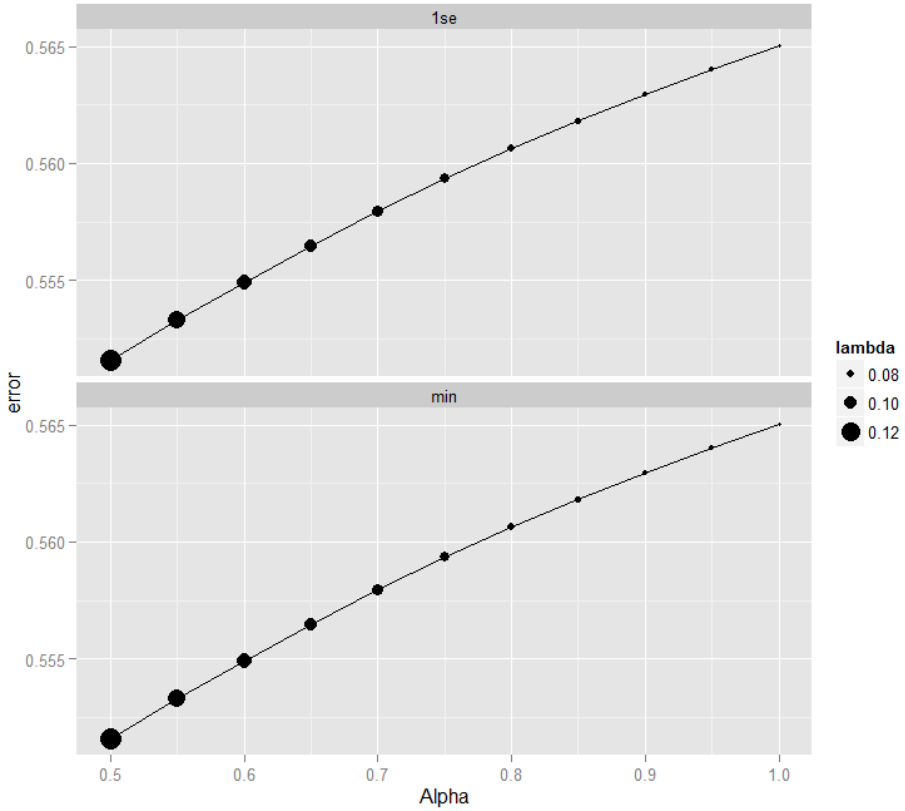

Figure 5.4: A plot of α vs cross-validation error. The lower the error the better. The top panel shows the error rate for the λ_{1se} approach ($\lambda_{1se} = 0.131$) and the bottom panel shows the error by the value of λ which minimises the error ($\lambda_{min} = 0.131$).

Knowing the optimal value of α we can fit the model. To reiterate, fitting the model using λ_{1se} we should obtain a model which is simpler but less accurate than we may otherwise obtain. The diagnostic plot for this model can be seen on the left-hand side of Figure 5.5 below. The values at the top of the plot indicate the number of predictors included in the model. Each point and the vertical lines above and below it shows the cross-validation error and corresponding confidence intervals for different values of *log(λ)*. Although not clear in either plot as in both instances they overlap, two vertical dashed lines appear on each plot. The leftmost dashed vertical line illustrates the value of λ where the cross-validation error is smallest while the right dashed vertical line indicates the next largest value of λ error within one standard error of the minimum. Again,

following the principle of parsimony we may prefer a simpler but less accurate model depending on the circumstances.

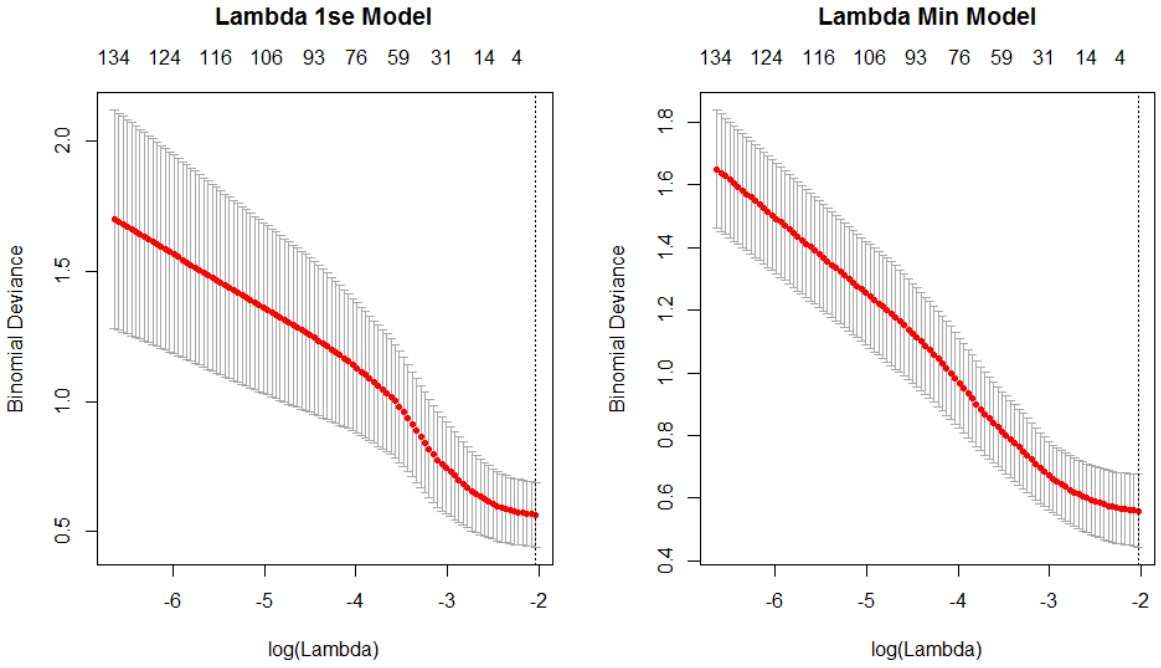

Figure 5.5: Diagnostic plot for the λ_{1se} and λ_{min} models for naturally-complete HEI data.

As the value of log(λ) increases, minimising the cross-validation error – that is the prediction errors made by each model as it is developed when tested on a subset of the data – becomes proportionately less important to the model development and minimising the number of metrics in the model becomes more important. In both the cases shown above as log(λ) increases the cross-validation error decreases regardless suggesting that a large number of the metrics have little if any predictive power when fitted using a cross-validated approach. Those metrics which remain in a model at each stage are those which minimise the sum of (1) the difference between the models predictions and the actual outcome, and (2) a penalty term which is reduced by having fewer metrics in the model.

The position of the overlapping dashed vertical lines in both the λ_{1se} (lefthand) and λ_{min} (righthand) plots indicates that, in both cases, the model with the lowest cross-validation error and the simpler model with an acceptably worse cross-validation error are both the same: they contain no metrics. Such a model suggests that no combination of the metrics provide a significantly better prediction of which providers are more likely to be judged 'unsatisfactory' than simply assuming all providers have an equal chance. The λ_{1se} and λ_{min} models both predict each providers probability of being judged 'unsatisfactory' as:

$$P(Unsatisfactory) = \frac{e^{-2.5288}}{1 + e^{-2.5288}} = 7.39\%$$

This is simply the number of 'unsatisfactory' reviews in the data sets divided by the total number of reviews (13 ÷ 176 = 0.0738636).

Having considered all naturally-complete metrics with a feasible link to quality assurance, not just in their absolute state but also modified to account for changes over time – both in percentage and absolute terms - no suitable model has been obtained. No robust model exists which is a better predictor than simply assuming all HEIs have an equal chance of receiving an 'unsatisfactory' review. Therefore, the answer to our first question:

> *Using only naturally-complete metrics, could the outcome of QAA HEI reviews have been successfully predicted?*

is no. Even with perfect hindsight using only naturally complete metrics we could not have successfully predicted the outcome of QAA HEI reviews. Prioritising the QAA's reviews of HEIs based on some combination of these metrics would fare no better than treating all HEIs equally regardless of the data.

5.3. Results – Imputed Data

With the use of statistical imputation and all comprehensive metrics, could the outcome of QAA HEI reviews have been successfully predicted?

If we are unable to reliably predict the outcome of QAA reviews when we discard metrics with missing values, can we have more success when we estimate these missing values and retain the metrics? Doing so provides us with 299 additional metrics and gives us a better chance of forming an accurate model.

5.3.1. Initial Data Exploration

As the imputed data set contains all the metrics that feature in the naturally-complete data set there is substantial overlap in the metrics with significant p-values. In total there were 42 metrics with a p-value less than 0.05 and 270 with a p-value less than 0.25, almost exactly in line with the number we would expect to occur by chance alone. Table 5.2 below shows the data sets from which the *p<0.05* metrics came:

Metric Type	Metrics with p-value < 0.05
Applications	11
Staffing	7
Students	7
HESA Performance Indicators	5
Research	3
UCAS	3
Finance	3
Destination of Leavers Survey	1
National Student Survey (NSS)	1
Unit Expenditure	1

Table 5.2: A breakdown of the metric types with a p-value less than 0.05 for the imputed HEI data set.

Once more, the majority of the metrics are not those one would likely select *a priori* to predict the outcome of QAA reviews. The one NSS metric is, as before, *NSS005_Abs - Q5 - The criteria used in marking have been clear in advance*. The one DLHE metric is *DLH012_Abs – The proportion of UK domiciled total leavers who obtained qualifications through full-time study and were reported as unemployed.* These metrics are, however, dwarfed in number by the applications, staffing, students and other indicators.

The metric with the lowest p-value: *APL004_Cp1 – the one-year percentage change in the proportion of successful applicants whose age is known who are aged 20 are & under* shown below in Figure 5.6:

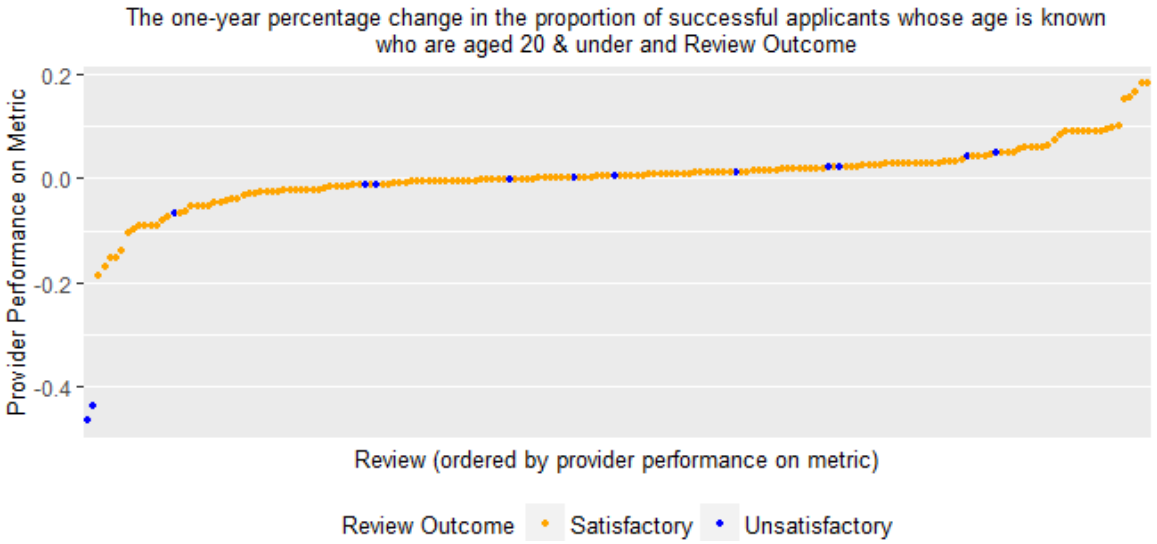

Figure 5.6: A plot of the latest value of 'the one-year percentage change in the proportion of successful applicants whose age is known who are aged 20 are & under' prior to each review and the outcome of that review.

Here we can see that the p-value is being strongly influenced by two outlying reviews which resulted in an 'unsatisfactory' rating following a decrease in the proportion of successful applicants who were aged 20 or under. Again the pattern does not hold for the remainder of the 'unsatisfactory' reviews which are evenly spread throughout a series of 'satisfactory' reviews. It is also once more the case that there is no obvious reason why there should be a link between the metric and the outcome of quality assurance reviews.

One of the more promising metrics on first inspection is *STA054_Abs - Proportion of staff (FTE) whose nationality is known who are of "Other-EU" nationality* shown below in Figure 5.7:

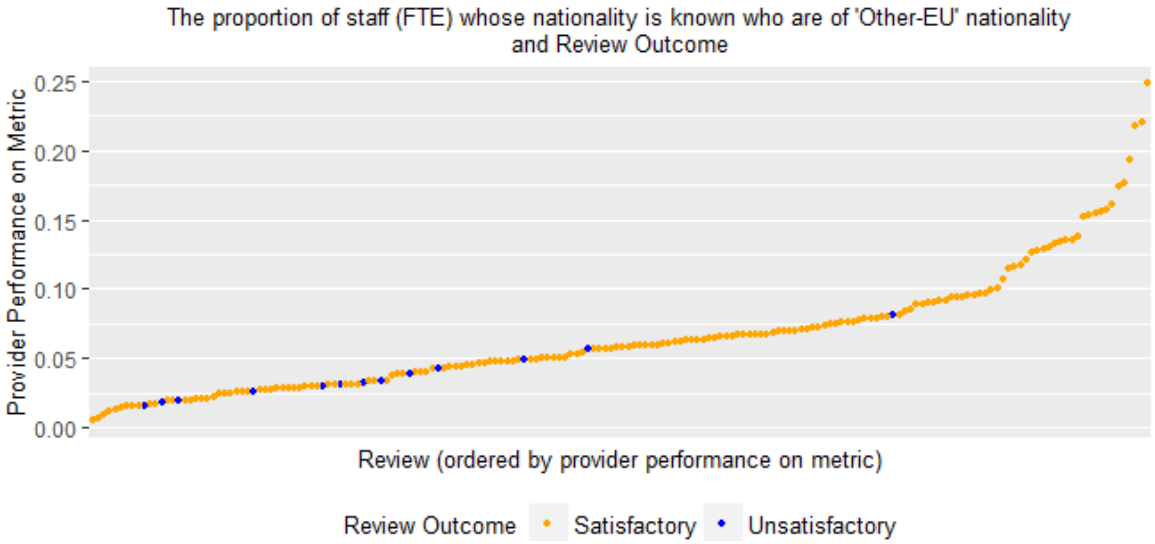

Figure 5.7: A plot of the latest value of '*Proportion of staff (FTE) whose nationality is known who are of "Other-EU" nationality*' prior to each review and the outcome of that review.

Those HEIs who received 'unsatisfactory' reviews tended to have a low proportion of EU staff who were not from the UK. As with all the other metrics discussed above however, these 'unsatisfactory' reviews share the same characteristics as a large number of 'satisfactory' reviews. Again, it is not immediately apparent why – although that is not to say a relationship is not meaningful – the proportion of FTE staff who are from the EU but are not British would impact on quality assurance review findings.

The fact that some of the metrics show a tendency for 'unsatisfactory' reviews to be situated at one end of the distribution in overall performance, as illustrated above, may be of limited use. If, as we see, the 'unsatisfactory' reviews are grouped with a larger number of 'satisfactory' reviews then the predictive power of that metric is extremely limited. Even if all of the 'unsatisfactory' reviews were preceded by performance in the worst third for a metric or collection of metrics for all reviews that would still mean there was only a 13 in 59 or 22% chance that an HEI in that bottom third of performance would subsequently be judged 'unsatisfactory'. It is unlikely that HEIs would

tolerate being prioritised for review based on the results of a metric which only correctly predicted 'unsatisfactory' performance 22% of the time (or perhaps more aptly, incorrectly predicted 'unsatisfactory' performance 78% of the time). What will be of value is if a model can be created which combines the metrics of poor performance in a number of areas into a definitive signal that an HEI is likely to be 'unsatisfactory' whilst not prioritising too many HEIs performing poorly on some metrics but are 'satisfactory' overall.

5.3.2. Fitting the Model

With the imputed metrics we obtain the diagnostic plots shown below in Figure 5.8. As before using the λ_{1se} approach results in the intercept-only model which suggests all HEIs have an equal chance of receiving an 'unsatisfactory' review regardless of all the available data. With the inclusion of the imputed metrics however we obtain a λ_{min} model which contains three metrics.

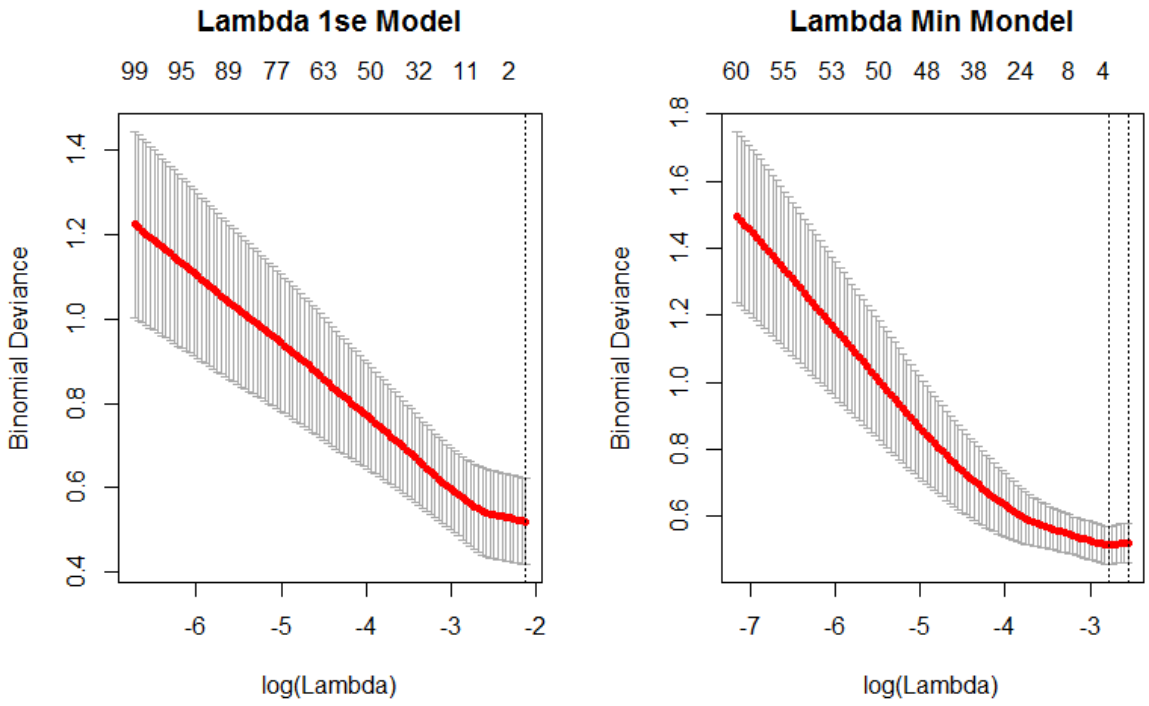

Figure 5.8: Diagnostic plots for the λ_{1se} (left) and λ_{min} (right) model for imputed HEI data.

The three metrics are:

1. APL006_Ca1 - the one-year change in the proportion of successful applicants whose age is known who are aged 25 & above.
2. KFI020_Abs - the percentage ratio of contribution from research grants & contracts to research grants & contracts income. This somewhat confusingly named metric is calculated as:

$$\frac{\text{Research grants \& contracts Income} - \text{Total research grants \& contracts expenditure}}{\text{Research grants \& contracts Income}}$$

and is a measure of the difference between how much money an HEI has received for, and spent on, research in a given year. High negative values indicate an HEI spending far more on research than the income it has received for that purpose in that year.

3. STA062_Ca1 - the one-year change in the proportion of full-time equivalent (FTE) staff who are principally financed by the institution.

The model calculates the probability of an HEI receiving an 'unsatisfactory' review using these three metrics as follows:

$$P(Uns) = \frac{e^{(-2.61 + (5\times APL006\_Ca1) + (-0.000088\times KFI020\_Abs) + (11.16\times STA062\_Ca1)}}{1 + e^{(-2.61 + (5\times APL006\_Ca1) + (-0.000088\times KFI020\_Abs) + (11.16\times STA062\_Ca1)}}$$

As the coefficient (the number by which the metric is multiplied in the above equation) is positive for the applications and staffing metrics , positive values of each metric lead to increases in the predicted probability of an 'unsatisfactory' review whilst negative values lead to decreases. Therefore, *ceteris paribus*, an increase in the proportion of successful applicants whose age is known who are aged 25 & above and the proportion of staff who are principally financed by the institution will lead to an increase in the predicted likelihood of being judged 'unsatisfactory'. As the coefficient is negative for the finance metric the opposite is true: negative values, indicating an HEI spending more on research than the income it has received for that purpose in a given year, will lead to increases in the predicted probability of an 'unsatisfactory' review whilst positive values will lead to decreases.

To demonstrate the effect that changes in performance for each metric will have on the predicted likelihoods of being judged 'unsatisfactory', the model is explored below for three hyopthetical HEIs.

An HEI with a decrease in the proportion of successful applicants whose age is known and who are aged 25 & above on the previous year will have a negative value of APL006_Ca1 which, when multiplied by the positive coefficient (5.00), will decrease the predicted likelihood of being judged 'unsatisfactory'. If we consider three hypothetical HEIs each of which previously had 10% of their succesful applicants whose age was known who were over 25, and finance and staffing numbers which remained constant such that KFI020_Abs and STA062_Ca1 equal zero, then we can

demonstrate the effects of a change in the make-up of successful applicants. Let HEIs A, B and C respectively see their proportion of succesful applicants whose age is known and who are aged 25 or over this year be 20%, 10% and 5% respecitvely. This is a 100% increase, no change, and a 50% decrease:

HEI	Previous Proportion	Current Proportion	APL006_Ca1	Probability of being judged 'unsatisfactory'
A	0.10	0.20	0.20 – 0.10 = 0.10	$\frac{e^{(-2.61 + (5.00 \times 0.1) + 0 + 0)}}{1 + e^{(-2.61 + (5.00 \times 0.1) + 0 + 0)}} = 10.813\%$
B	0.10	0.10	0.10 – 0.10 = 0	$\frac{e^{(-2.61 + (5.00 \times 0) + 0 + 0)}}{1 + e^{(-2.61 + (5.00 \times 0) + 0 + 0)}} = 6.845\%$
C	0.10	0.05	0.05 – 0.10 = -0.05	$\frac{e^{(-2.61 + (5.00 \times -0.05) + 0 + 0)}}{1 + e^{(-2.61 + (5.00 \times -0.05) + 0 + 0)}} = 5.417\%$

Table 5.3: The hypothetical application outcomes of three HEIs and the resulting predicted likelihood of being judged 'unsatisfactory'.

HEI A having doubled its proportion of succesful applicants aged 25 or over has a predicted likelihood of being judged 'unsatisfactory' of 10.81% whereas HEI C, which has halved its proportion of succesful applicants aged 25 and over, has a predicted likelihood of being judged 'unsatisfactory' of 5.42%. Whilst the probability of either being judged 'unsatisfactory' remains low in absolute terms, HEI A is now almost twice as likely to be judged 'unsatisfactory' than HEI C.

An HEI spending more money on research than it receives specifically for that purpose will have a negative value of KFI020_Abs which, when multiplied by the negative coefficient (-0.000088), will increase the predicted likelihood of being judged 'unsatisfactory'. If we again consider three hypothetical HEIs each of which have received £10M in research funding, and have maintained staffing and application numbers such that STA035_Ca1 and APL006_Ca1 equal zero, then we can demonstrate the effects of research spending levels on the predicted likelihood of being 'unsatisfactory'. Let HEI A spend £20M on research, HEI B £10M on research, and HEI C £5M on research. This is twice as much, exactly the same, and half what they are respectively funded specifically for this purpose:

HEI	Research Funding	Research Expenditure	KFI020_Abs	Probability of being judged 'unsatisfactory'
A	£10M	£15M	$\frac{10-20}{10} = -1$	$\frac{e^{(-2.61 + 0 + (-0.000088 \times -1) + 0)}}{1 + e^{(-2.61 + 0 + (-0.000088 \times -1) + 0)}} = 6.8503\%$
B	£10M	£10M	$\frac{10-10}{10} = 0$	$\frac{e^{(-2.61 + 0 + (-0.000088 \times 0) + 0)}}{1 + e^{(-2.61 + 0 + (-0.000088 \times 0) + 0)}} = 6.84976\%$
C	£10M	£5M	$\frac{10-5}{10} = 0.5$	$\frac{e^{(-2.61 + 0 + (-0.000088 \times 0.5) + 0)}}{1 + e^{(-2.61 + 0 + (-0.000088 \times 0.5) + 0)}} = 6.84948\%$

Table 5.4: The hypothetical research funding and expenditure of three HEIs and the resulting predicted likelihood of being judged 'unsatisfactory'.

This time the effect is clearly more marginal. HEI A overspending by 100% has a predicted likelihood of being judged 'unsatisfactory' of 6.85% whereas HEI C which spends just half its budget has a minutely-lower predicted likelihood of being judged 'unsatisfactory' of 6.849%. It is important to note that this metric deals with relative amounts and therefore sees no difference between an HEI which spends £2bn on reaseach when only allocated £1bn, and an HEI which spends £2,000 on research when it has only been allocated £1,000.

Finally, an HEI which increases the proportion of full-time equivalent staff it prinicipally finances will have a positive value of STA062_Ca1 which, when multiplied by the positive coefficient (11.16), will increase the predicted likelihood of being judged 'unsatisfactory'. If we again consider three hypothetical HEIs each of which had 5% full-time equivalent staff on permanent or open-ended contracts in the previous year, and have maintained applicant and finance numbers such that APL006_Ca1 and KFI020_Abs equal zero, then we can demonstrate the effects of staffing levels on the predicted likelihood of being 'unsatisfactory'. Let HEI A principally finance 10% of its full-time equivalent staff, and HEIs B and C 5% and 2.5% respectively. This is double, exactly the same, and half the proportion each respective instiution had the year before:

HEI	Proportion last year	Proportion this year	STA062_Ca1	Probability of being judged 'unsatisfactory'
A	5%	10%	0.10 – 0.05 = 0.05	$\frac{e^{(-2.61 + 0 + 0 + (11.16 \times 0.05))}}{1 + e^{(-2.61 + 0 + 0 + (11.16 \times 0.05))}} = 11.385\%$
B	5%	5%	0.05 – 0.05 = 0	$\frac{e^{(-2.61 + 0 + 0 + (11.16 \times 0))}}{1 + e^{(-2.61 + 0 + 0 + (11.16 \times 0))}} = 6.85\%$
C	5%	2.5%	0.025 – 0.05 = -0.025	$\frac{e^{(-2.61 + 0 + 0 + (11.16 \times -0.025))}}{1 + e^{(-2.61 + 0 + 0 + (11.16 \times -0.025))}} = 5.270\%$

Table 5.5: The hypothetical proportions of staff funded principally by the institution for three HEIs and the resulting predicted likelihood of being judged 'unsatisfactory'.

The effects are similar to those seen for metric *APL006_Ca1* in table 5.3. With an 11.385% probability of being judged 'unsatisfactory' HEI A is roughly twice as likely as HEI C to be 'unsatisfactory'; however, in absolute terms the likelihood of either being judged 'unsatisfactory' remains low.

As shown in Table 5.6 below, if we combine these three metrics we can see that HEI A which has doubled its proportion of succesful applicants aged 25 or over, is spending 50% more on research than it had received for that purpose, and has doubled the proportion of full-time equivalent staff it principally employed has a 17.48% predicted likelihood, or roughly 1 in 6 chance, of being judged 'unsatisfactory'. HEI C on the other hand, spending only half of its research funding budget and having halved the number of staff principally financed by the institution, has just a 4.15% predicted likelihood, or roughly 1 in 24 chance, of being judged 'unsatisfactory'.

HEI	Probability of being judged 'unsatisfactory'
A	$\frac{e^{(-2.61+(5\times0.1)+(-0.000088\times-1)+(11.16\times0.05))}}{1+e^{(-2.61+(5\times0.1)+(-0.000088\times-1)+(11.16\times0.05))}} = 17.48103\%$
B	$\frac{e^{(-2.61+(5\times0)+(-0.000088\times0)+(11.16\times0))}}{1+e^{(-2.61+(5\times0)+(-0.000088\times0)+(11.16\times0))}} = 6.84976\%$
C	$\frac{e^{(-2.61+(5\times-0.05)+(-0.000088\times0.5)+(11.16\times-0.025))}}{1+e^{(-2.61+(5\times-0.05)+(-0.000088\times0.5)+(11.16\times-0.025))}} = 4.152515\%$

Table 5.6: The hypothetical values of APL006_Ca1, KFI020_Abs and STA062_Ca1 metrics and the resulting predicted likelihood of an HEI being judged 'unsatisfactory'.

As with the metrics considered in isolation, it is clear that the effects of changes in performance on the predicted likelihood of being judged 'unsatisfactory' can be substantial in relative terms, but are still fairly limited in absolute terms. Extreme performance still results in a predicted likelihood of being 'satisfactory' greater than 80%. The narrow range of predicted probabilities suggest that the model is unlikely to perform well. Neither the highest nor lowest probabilities are far from the proportion of reviews which were judged unsatisfactory in the data set (13 ÷ 184 = 7.07%) which was the sole predictor used in the λ_{1se} model. The limited range of predicted probabilities is the result of a lack of certainty in the model; when a model has difficulty identifying which HEIs will or will not be found 'unsatisfactory' the predicted likelihoods of the event occuring will always be clustered close to the probability of being judged 'unsatisfactory' regardless of the data.

5.3.3. Evaluating the Model

Figure 5.9 below shows the ROC curve for this model which has a fairly weak 'area under the curve' value of 0.720 suggesting a poor rate of 'unsatisfactory' HEIs being succesfully prioritised as the threshold criterion for triggering a review is lowered:

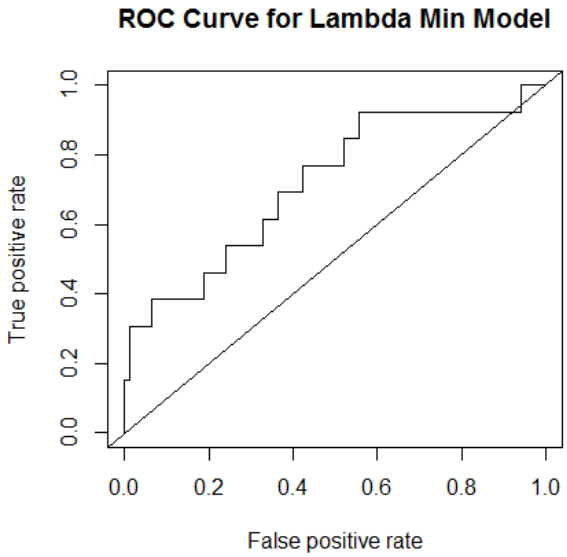

Figure 5.9: The ROC curve for the HEI model featuring the imputed metrics APL006_Ca1, KFI020_Abs and STA062_Ca1.

Table 5.7 below shows in greater detail the effect of lowering the threshold required for the model's predicted probability of failure to trigger a review. The first four 'unsatisfactory' reviews could have been predicted with only two 'satisfactory' reviews being incorrectly prompted; however, the predictions worsen after intial success.

	Predicted Probability of an 'unsatisfactory' outcome required to trigger a review	Number of 'unsatisfactory' reviews (true positives)	Number of 'satisfactory' reviews (false positives)	Accuracy rate
← Decreasing probability of a review being 'unsatisfactory' required to trigger inspection	23.09%	1	0	0.93
	19.76%	2	0	0.94
	9.68%	3	2	0.93
	9.45%	4	2	0.94
	8.17%	5	11	0.90
	7.21%	6	32	0.79
	7.10%	7	41	0.74
	7.01%	8	56	0.67
	6.94%	9	62	0.64
	6.89%	10	72	0.59
	6.81%	11	89	0.51
	6.74%	12	95	0.48
	5.91%	13	161	0.13

Table 5.7: The number of 'satisfactory' and 'unsatisfactory' HEI reviews that would have resulted from decreasing the threshold required to prompt a review based upon the λ_{min} model.

To have detected all the 'unsatisfactory' reviews the QAA would have had to conduct 161 additional 'satisfactory' reviews. Therefore, 174 reviews would have been required to detect the 13 'unsatisfactory' providers – an error rate of 92.53%. Even if the one 'hard to predict', outlying 'unsatisfactory' provider is removed from the equation, 95 unnecessary reviews would have to be conducted to discover the 12 cases of 'unsatisfactory' performance – an error rate of just 88.79%. The accuracy rate detailed in Table 5.7 – the proportion of correct predictions made by the model – is deceptive as there is a high proportion of HEIs that are 'satisfactory'; simply predicting all HEIs will be 'satisfactory' would result in a reasonable accuracy rate of 171 ÷ 184 = 92.9%. The model's performance quickly falls below this level after its intitial success.

Figure 5.10 below shows the same data as Table 5.7 but graphically. It is clear there is one 'unsatisfactory' review which the model struggles to predict and the remainder are spread throughout a large number of 'satisfactory' reviews.

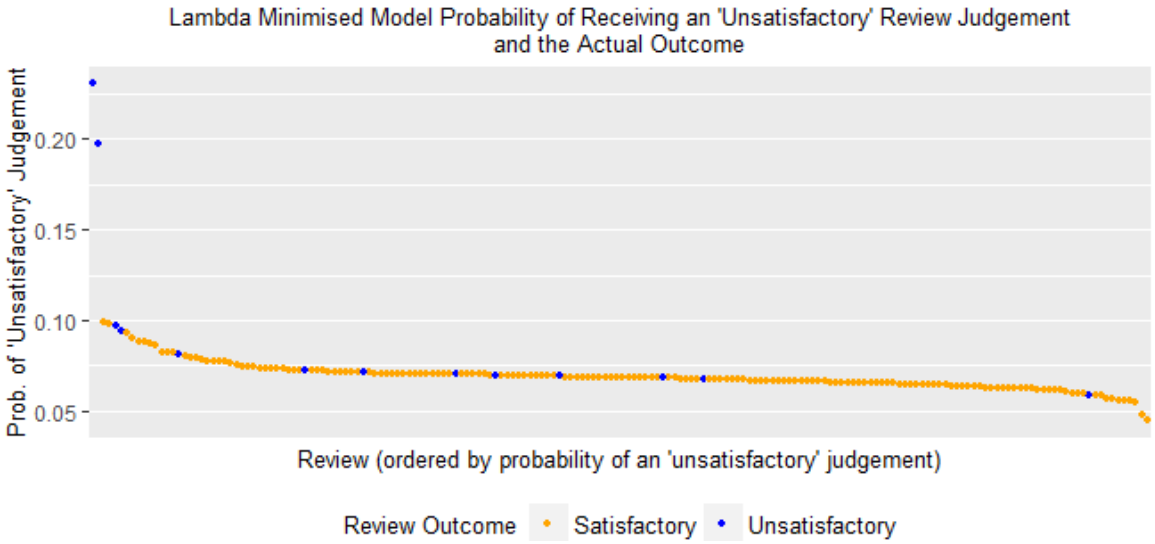

Figure 5.10: Probabilities predicted by the imputed-data model of each of the 184 complete, comparable HEI reviews and their actual outcome.

As shown with the worked example of the model above, the narrow range of predicted probabilities does not mean small changes in metric performance lead to significant changes in predicted probabilities. Rather, the lack of relationship between the metrics and the outcome of QAA reviews seemingly results in the best possible model struggling to predict an 'unsatisfactory' outcome with any degree of certainty.

As stated in the previous methods chapter, witholding data to test the model or waiting for new reviews to take place is not an option for HEIs. Instead, two alternative approaches are adopted to sense check the model. Figure 5.11 below show the model's application to the 2012/13 data along with each HEIs (somewhat contested) Guardian rankings for that academic year. This gives

us an approximation of which HEIs would have been prioritised for review in 2014/15 had the model been in place. There is still a strong clustering of highly-ranked HEIs predicted as being most likely to have an 'unsatisfactory' review including three of the UK's ostensibly top 10 unviersities in the 25 most likely to be 'unsatisfactory'. Of greater concern however is the even narrower range of predicted probabilities. The model once again struggles to predict 'unsatisfactory' reviews with any certainty as the underlying relations are so weak. Putting aside the narrow range of probabilities, some highly-ranked HEIs are considered two or three times more likely to be 'unsatisfactory' than many lower ranked HEIs. There are two possible non-exclusive explanations for this: the model is nonsensical and/or there are a number of highly-regarded HEIs in the UK who have significant quality assurance issues which are going unnoticed.

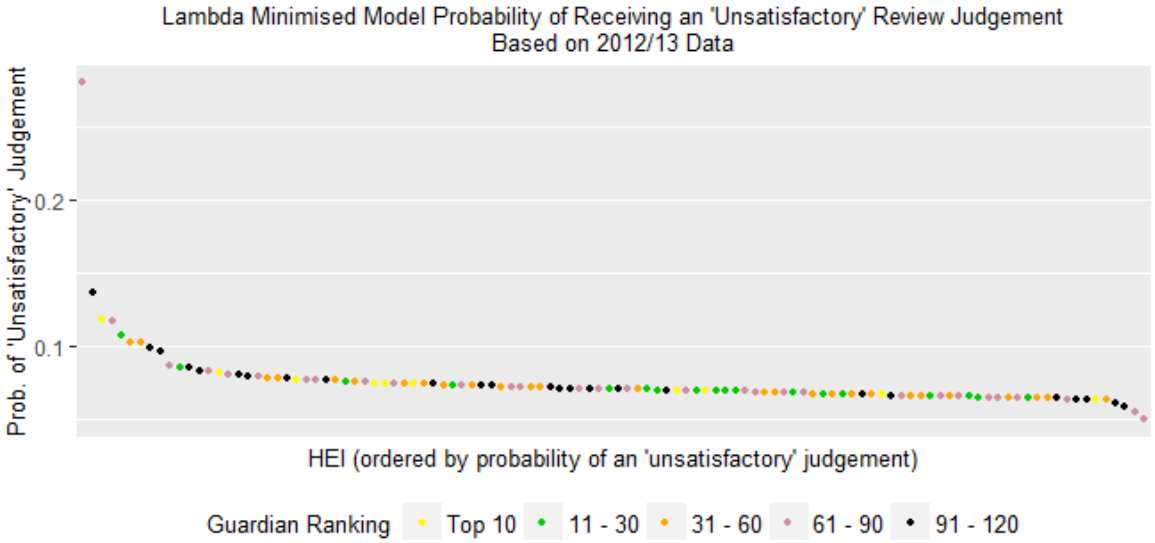

Figure 5.11: Probabilities predicted by the HEI model featuring the imputed metrics APL006_Ca1,KFI020_Abs and STA062_Ca1 of each UK HEI receiving an unsatisfactory review based on 2012/13 data.

Taking an historic point in time, in this case 1st October 2009, and using the most up-to-date instances of the three metrics in the model available at the time we can gain a snapshot of the prioritisation list the QAA would have had at that time. Those HEIs that were reviewed within one year of 1st October 2009 have the outcome of their review shown in Figure 5.12 below.

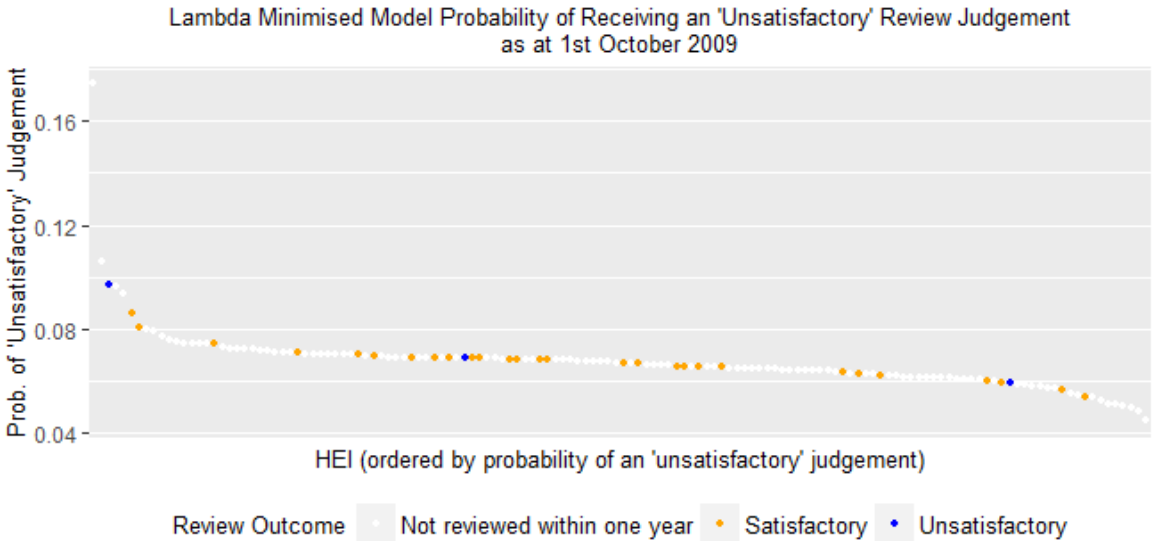

Figure 5.12: Probabilities predicted by the HEI model featuring the imputed metrics APL006_Ca1,KFI020_Abs and STA062_Ca1 of each UK HEI receiving an unsatisfactory review based on latest available data on 1st October 2009.

We can see that Brunel University, judged to be 'unsatisfactory' before the end of 2009, was ranked as one of the least likely to HEIs to be 'unsatisfactory'. Of the remaining two HEIs judged 'unsatisfactory' within one year one was reassuringly ranked third most likely to be; however the other would have only been prioritised behind 49 others. Again, the range of predicted probabilities is very narrow and we can either reason that the model is very poor and would have led to a very high number of incorrectly prioritsied reviews or, less likely, there were very many 'unsatisfactory' HEIs which were not reviewed.

5.3.4. Summary

For this question we have considered all imputed metrics with a feasible link to quality assurance that could form part of a cost-effective, data-driven, risk-based approach, not just in their absolute state but also modified to account for changes over time, both in percentage and absolute terms. The addition of the imputed metrics has meant we have been able to develop a predictive model better than simply assuming an equal probability of all HEIs being 'unsatisfactory':

$$P(Uns) = \frac{e^{(-2.61 + (5\times APL006\_Ca1) + (-0.000088\times KFI020\_Abs) + (11.16\times STA062\_Ca1)}}{1 + e^{(-2.61 + (5\times APL006\_Ca1) + (-0.000088\times KFI020\_Abs) + (11.16\times STA062\_Ca1)}}$$

However, the answer to question two:

> *With the use of statistical imputation and all comprehensive metrics, could the outcome of QAA HEI reviews have been successfully predicted?*

is no. The very narrow range of predicted probabilities contained in the best possible model shows that there is little relation between the metrics and the outcome of QAA reviews. This is reflected

in the fact that for all of the 13 'unsatisfactory' reviews to have been prioritised (even operating with the benefit of perfect hindsight), 161 unnecessary 'satisfactory' reviews would have to have been conducted. This is only ten reviews short of having conducted all the reviews anyway and represents a 92.5% error rate amongst the HEIs reviewed. Furthermore, examining how the model would have performed at a specific point in time shows a similar picture; 121 HEIs were predicted as more likely to be 'unsatisfactory' on 1st October 2009 than Brunel, found 'unsatisfactory' just ten weeks later. Finally, looking at the latest available data at the time of writing the predictions are again a cause for concern with three of the UK's top 10 HEIs in the top 25 most likely to be 'unsatisfactory'.

5.4. Results – Imputed Data Standardised In-Year

With the use of statistical imputation and all in-year standardised, comprehensive metrics, could the outcome of QAA HEI reviews have been successfully predicted?

If statistical imputation has provided us with a better model than the one based on only the naturally complete data, can standardising this data within each year further improve the model by accounting for sector-wide trends over time in the data?

5.4.1. Initial Data Exploration

The standardised dataset is based on the imputed data set and therefore contains the exact same number of metrics. We would therefore expect to see a similar number of metrics with a p-value of less than 0.05 and this was the case. Again there were 42 metrics with a p-value less than 0.05 although there was some difference in the metrics. The one previously significant DLHE metric no longer features suggesting changes in employment rates over time were having a confounding effect. Conversely, the student:staff ratio becomes a significant univariate metric at the 5% level once general trends over time are accounted for. Table 5.8 below shows the data sets from which the metrics with a p-value less than 0.05 came:

Data Set	Number of metrics where $p<0.05$
Applications / UCAS	14
Students	9
Finance	9
HESA Performance Indicators	4
Research Statistics	3
Staffing	2
Student:Staff Ratio	1

Table 5.8: A breakdown of the metrics with a p-value less than 0.05 for the imputed, standardised HEI data set.

The standardisation of the metrics has led to an increase in the already proportionately high number of metrics relating to applications which have a significant relationship with the outcome of the QAA reviews. The distribution of APL025_Ca2 – *the two-year change in the total number of applicants,* the metric with the lowest p-value, shown below in Figure 5.13, indicates a relationship between institutions receiving a greater number of applications prior to their review than two years before it.

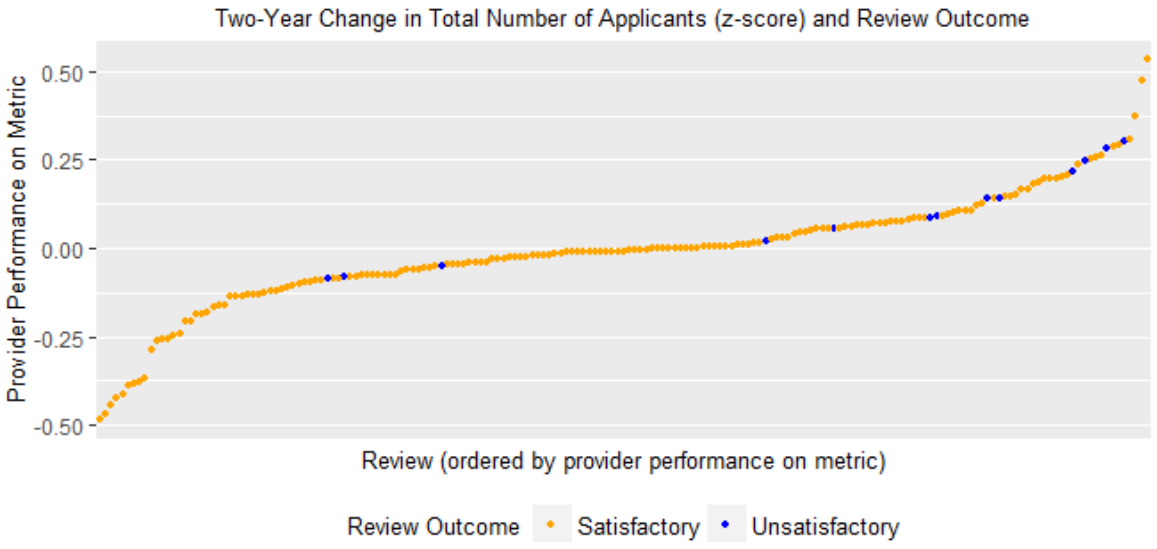

Figure 5.13: A plot of the two-year change in the standardised total number of applicants prior to each review and the outcome of that review.

Again, we see the familiar pattern of more than half of the 'unsatisfactory' providers towards one extreme of the distribution distributed amongst a considerable number of 'satisfactory' providers and several 'unsatisfactory' providers towards the other end of the distribution. One can see an intuitive link that providers with declining quality, which has not been prevented due to 'unsatisfactory' quality assurance, could suffer a decline in the proportion of applications compared to other HEIs.

One other metric of note to have a p-value less than 0.05 with the use of in-year standardisation was the student:staff ratio calculated simply as the number of students at an institution divided by the number of staff. When a metric is standardised the mean is subtracted from a value so those institutions with a negative value on the y-axis of Figure 5.14 below had a below average number of students per member of staff. It is this characteristic – having fewer students per member of staff than average – which somewhat counterintuitively correlates with being judged 'unsatisfactory'.

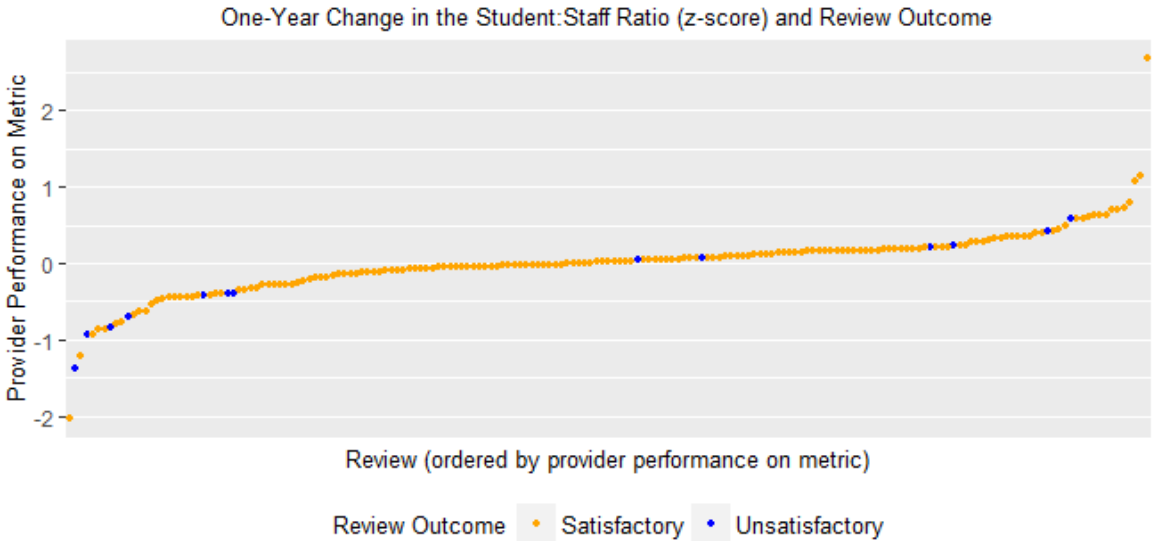

Figure 5.14: A plot of the one-year change in the in-year standardised student to staff ratio prior to each review and the outcome of that review.

The next step is to determine whether any combination of these metrics could have provided an effective, cross-validated model which would have allowed for the 'unsatisfactory' reviews to have been successfully prioritised.

5.4.2. Fitting the Model

Figure 5.15 below shows that, as with the naturally-complete data, no cross-validated model provides a better fit than simply assuming all institutions have the same probability of being judged 'unsatisfactory' regardless of the metrics available:

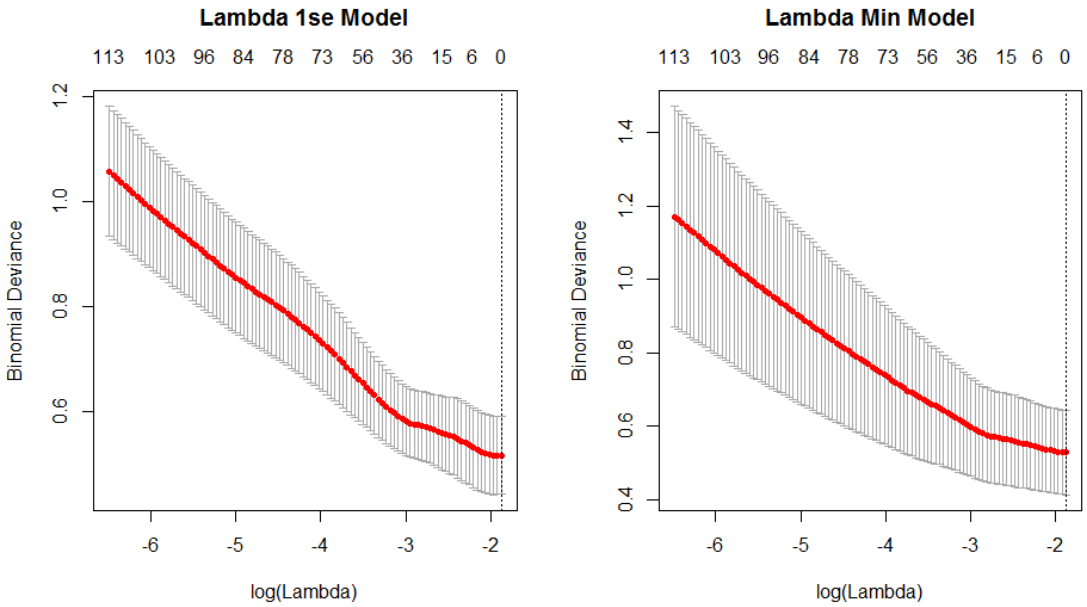

Figure 5.15: The diagnostic plots for the λ_{1se} (left) and λ_{min} (right) model for standardised, imputed HEI data.

Therefore, the answer to our question:

> *With the use of statistical imputation and all in-year standardised, comprehensive metrics, could the outcome of QAA HEI reviews have been successfully predicted?*

is again, no. The obvious question at this point is how is this the case when the non-standardised imputed data did produce a model, albeit a questionable one? The answer is that for HEIs, absolute performance is seemingly a better predictor than relative. For example, if a single HEI were in £5,000,000 worth of debt in 2008 whilst all other HEIs were debt free, it is still in a better position than being in £75,000,000 worth of debt in 2012 when all other HEIs are in £100,000,000 worth of debt. Despite being relatively worse off than other HEIs in 2008 and relatively better off than other HEIs in 2012, it is still the case that the HEI's financial position is far worse in 2012, and likely putting a far greater strain on quality assurance activities. One HEI's debt does not affect the ability of another's to fund quality assurance or other activities. Clearly this may not be the case for all metrics, but it appears to be the case for those which give the best prediction of the outcome of QAA reviews.

5.5. Results – Benchmarked Data

With the use of statistical imputation and all comprehensive, benchmarked metrics, could the outcome of QAA HEI reviews have been successfully predicted?

If standardising the data by year has failed to provide an adequate model, might benchmarking the data – standardising it by HEI type – prove more successful?

5.5.1. Initial Data Exploration

Coincidentally, there was the same number of metrics with a p-value less than 0.05 for the benchmarked data as there was for the standardised data set, despite the benchmarked data set containing a larger number of metrics. The 42 metrics with a p-value less than 0.05 were markedly different from the standardised data set; the 'Applications/UCAS' data which contributed the greatest number of significant individual metrics for the standardised data set does not feature at all for the benchmarked data set. Instead, staffing, students and finance metrics are more significant. The one NSS measure significant in the earlier analyses is no longer present. Table 5.9 below shows the specific data sets from which the metrics with a p-value less than 0.05 came:

Data Set	Number of metrics where *p<0.05*
Finance	13
Students	12
Staffing	9
Research	3
HESA Performance Indicators	2
Destination of Leavers	1
Previous Review Findings	1

Table 5.9: A breakdown of the metrics with a p-value less than 0.05 for the imputed, benchmarked HEI data set.

The distribution of *UFI086_Abs – the retained proceeds of sales (total capital expenditure),* the metric with the lowest p-value shown below in Figure 5.16, indicates a relationship between institutions receiving a greater amount of funds retained from sales of capital than their benchmark peers and the outcome of QAA reviews.

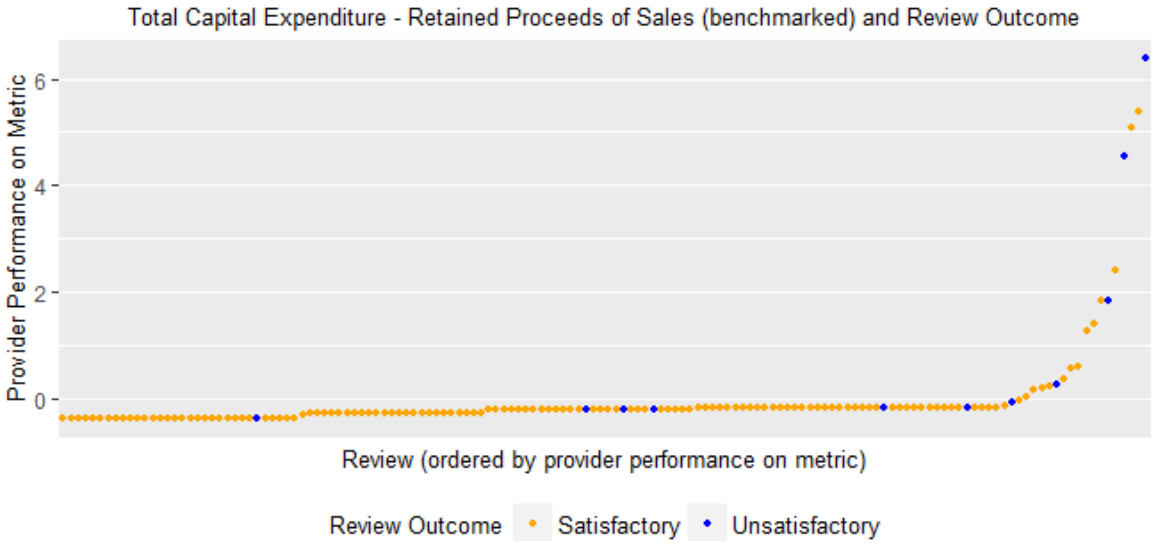

Figure 5.16: A plot of the benchmarked 'retained proceeds of sales' recorded under total capital expenditure prior to each review and the outcome of that review.

Again, there is the familiar pattern of several 'unsatisfactory' providers with extreme values at one end of the distribution, but the remainder distributed evenly throughout the 'satisfactory' providers. It is not immediately obvious what causal link could exist between the metric score and review findings.

Figure 5.17 below shows a similar pattern with two 'unsatisfactory' providers having an extremely high proportion of their academic staff who are leaving being both 'teaching and research' staff compared to their peers. The remaining 'unsatisfactory' providers are however evenly distributed throughout the 'satisfactory' providers. One can envision the proportion of staff who are leaving being 'teaching and research' staff impacting on quality assurance activities, but the proportion of staff is not the same as number of staff (it could be minimal), and more staff focused on teaching only could be being hired to replace them.

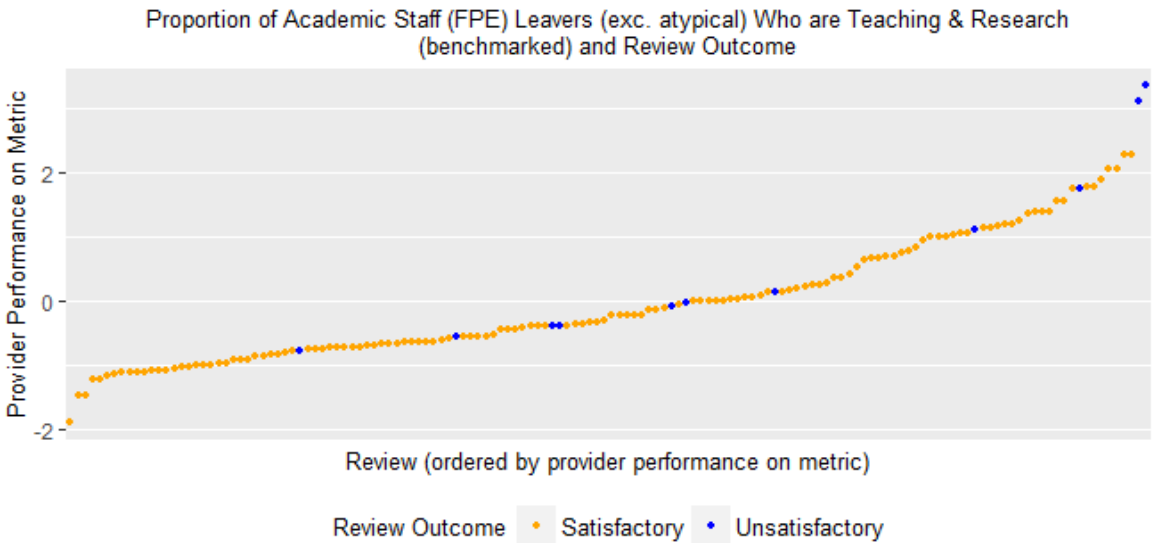

Figure 5.17: A plot of the benchmarked proportion of full-person equivalent academic staff leavers (excluding atypical leavers) who were classed as 'teaching & research' prior to each review and the outcome of that review.

The next step is to determine whether any combination of benchmarked metrics could have provided an effective, cross-validated model which would have allowed for the 'unsatisfactory' reviews to have been successfully prioritised.

5.5.2. Fitting the Model

Figure 5.18 below shows that, as with the naturally-complete data and in-year standardised data, no cross-validated model provides a better fit than simply assuming all institutions have the same probability of being judged 'unsatisfactory' regardless of the metrics available:

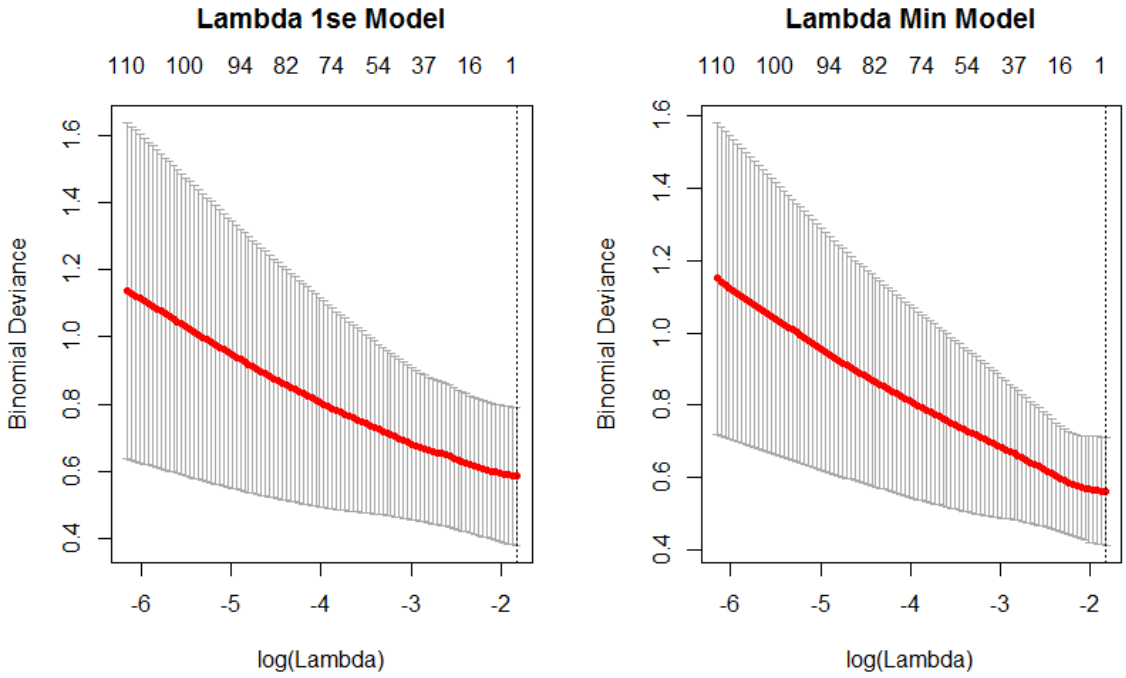

Figure 5.18: The diagnostic plots for the λ_{1se} (left) and λ_{min} (right) model for benchmarked, imputed data.

Therefore, the answer to the question:

> *With the use of statistical imputation and all comprehensive, benchmarked metrics, could the outcome of QAA HEI reviews have been successfully predicted?*

is again, no. Considering metric performance in relation to similar HEIs, rather than all HEIs with their differing mission, has failed to produce a model which is any better than ignoring the data and assuming all HEIs have an equal chance of being judged 'unsatisfactory'. This finding adds further weight to the suggestion in 5.4.2 above that absolute metric performance has a stronger relation with the outcome of QAA reviews than relative performance. Just as increasing the

amount of debt an HEI is in worsens its financial position and may affect its ability to fund quality assurance activities, regardless of whether other HEI's take on even more debt at the same time, so too will taking on more debt regardless of whether an HEI's peers have taken on even more debt.

5.6. Summary

Having analysed over 750 metrics, all those that could feasibly form part of a cost-effective, data-driven, risk-based approach, not just in their absolute form but multiple change-over-time variants, in their natural, imputed, standardised and benchmarked form, no possible combination of metrics could have effectively predicted the outcome of previous QAA reviews of HEIs, even with perfect hindsight.

The best model we were able to obtain was based on the imputed data:

$$P(Uns) = \frac{e^{(-2.61 + (5\times APL006\_Ca1) + (-0.000088\times KFI020\_Abs) + (11.16\times STA062\_Ca1)}}{1 + e^{(-2.61 + (5\times APL006\_Ca1) + (-0.000088\times KFI020\_Abs) + (11.16\times STA062\_Ca1)}}$$

However, there are significant concerns about the effectiveness of this model. Most notably, the very narrow range of predicted probabilities which resulted in 'unsatisfactory' reviews being distributed throughout 'satisfactory' reviews when ordered by the predicted likelihood of being 'unsatisfactory', and a very high error rate (92.5%). When applied to the 2012/13 data, the range of predicted probabilities remained very narrow and a large proportion of well-ranked HEIs were amongst those predicted as most likely to be judged 'unsatisfactory'. This is in part a result of the model making ineffective predictions, but it may also be the case that there is a significant amount of 'unsatisfactory' quality assurance being conducted in the UK's purported top HEIs. Finally, when applied to the data available at a fixed historical point in time to provide a realistic view of what the QAA would have been faced with, the model placed two of the three HEIs found 'unsatisfactory' within one year a long way down the prioritisation list.

Evaluation of the best possible model strongly suggests there is no effective relationship between the metrics and the outcome of QAA reviews. Looking at the metrics contained in the model, this is perhaps unsurprising. It is not entirely apparent *why* they should be able to foretell quality assurance failures. Considering the three metrics in turn:

1. APL006_Ca1 - the one-year change in the proportion of successful applicants whose age is known who are aged 25 & above.

2. KFI020_Abs - the percentage ratio of contribution from research grants & contracts to research grants & contracts income.
3. STA062_Ca1 - the one-year change in the proportion of full-time equivalent (FTE) staff who are principally financed by the institution.

An increase in the value of the staffing metric could feasibly (although tenuously) serve as a proxy for flagging institutions who are changing their workforce which creates disruption and introduces a lack of continuity. There are many possible challenges to this explanation however. The metric is based on relative numbers and does not account for base values; for example, if two HEIs both had an equal number of students and one doubled its staff, principally financed by the institution, from 500 to 1,000 and the other doubled its staff from 5,000 to 10,000 they would both be predicted to be equally and highly likely to be judged 'unsatisfactory. This is despite the fact that the latter HEI would have ten times as many experienced staff familiar with the existing quality assurance processes and able to protect quality and standards.

An alternative explanation for an increase could be that an HEI is maintaining its staffing numbers but that these staff are not as able to attract research funding as they were in the past and so are now reliant on the institution. If so, this would show a similar outcome to KFI020_Abs; however, it is difficult to see how either relates to one or more of the four areas covered by a QAA review: *academic standards*, the quality of *teaching and learning*, the *provision of information* or *enhancement*. Likewise, it is difficult to fathom how an increase in the proportion of undergraduate students whose age is known and who are aged 25 or over could feasibly serve as an indicator of *academic standards*, the quality of *teaching and learning*, the *provision of information* or *enhancement*. Furthermore, the metric only accounts for relative, not absolute, changes and so is more sensitive to changes at smaller institutions.

The best model therefore not only fails to identify 'unsatisfactory' provision with any degree of certainty, has a very high error rate and a questionable output when applied to 1st October 2009 and the latest data, but it is also not intuitive either. There is no obvious reason why a model comprising these three metrics would have the ability to forecast the outcome of QAA reviews.

5.7. Discussion

A data-driven, risk-based approach to prioritising QAA HEI reviews, as envisioned in the 2011 White Paper *Students at the Heart of the System*, will not work. Despite considering all metrics that could feasibly form part of a practical risk-based approach, not just in their natural form but

calculating change-over-time, imputed, standardised and benchmarked variants, no effective model exists. One immediate question that results from this finding is what does this mean for higher education policy? If a data-driven, risk-based approach cannot work for HEIs which account for approximately 90% of students, then at best such an approach can only be successfully applied to the minority of HE provided by FECs and alternative providers.

In November 2015 the quantitative findings from this thesis were made public. Senior Managers from QAA, HEFCE, BIS, and various representative bodies, senior academics and journalists attended a seminar hosted by King's Policy Institute and the International Centre for University Policy Research. The results were also presented at the annual conferences of the European Quality Assurance Forum and Society for Research into Higher Education, and published in the *Times Higher Education* magazine (EQAF, 2015; SRHE, 2015; Havergal, 2015). Finally, the author gave oral evidence to a closed session of BIS's *Assessing Quality in Higher Education* inquiry as well as submitting written evidence (Griffiths, 2015). Four months after the presentation of these findings, HEFCE published its 'Revised Operating Model for Quality Assessment' which (somewhat pointedly) stated:

> "Throughout this process, the relevant funding body will remain mindful of the complexities involved in making judgements about a higher education provider's performance, and will recognise that data analysis and dialogue in these circumstances needs to be robust, sophisticated and nuanced. It is particularly important to note here that we are not advocating a crude metrics-driven approach, using data to predict providers that might or might not have received successful outcomes under previous quality assessment approaches. Rather, data is used as one source of information to inform a broader judgement supported where needed by suitably qualified and independent experts."
>
> (HEFCE, 2016c, 109)

Similarly, in January 2016, Ian Kimber, Director of Quality Development at QAA, wrote:

> "On the face of it, [the findings of this thesis] could call into question the application of metrics to identify areas of 'risk' on which to focus quality assurance effort. However, I would argue that the findings underline the need for a contextualised and nuanced approach to the use of metrics, both in terms of identifying potential quality risk and of assessing student outcomes and their link to teaching quality."
>
> (Kimber, 2015)

It is not possible to say with certainty whether expert interpretation will be successful without conducting an empirical evaluation, prioritising one group of providers by expert consideration and a second group by data-driven approaches alone, including reviews of providers not prioritised by either method to detect false negatives. As noted in section 3.5.2 however, the literature suggests it is likely such an approach will be unsuccessful, especially in the low-validity, 'noisy' setting that is QAA review prioritisation. The difficulties of expert interpretation may be compounded by the lack of relation between indicators selected *a priori* by HEFCE, including NSS and continuation rate metrics, and the outcome of QAA reviews.

The issue of expert interpretation is somewhat secondary however. This analysis sought to determine which metrics could have successfully predicted the outcome of QAA HEI reviews, and with that whether or not a purely data-driven approach such as that envisaged in *Students at the Heart of the System* could be successful. It cannot, and this appears to have been accepted by the sector which has retreated from talk of such an approach to focus on a 'data informed' method instead.

HEIs are more data rich than FECs and alternative providers, but they are also very different. A purely data-driven, risk-based approach cannot work for HEIs, the next question is can such an approach work for FECs?

Appendix E – HEI Metrics

The set of 754 metrics used in this study prior to change-over-time, in-year standardisation and benchmarking calculations being added.

Area	Metric Code	Proposed metrics
Applications	APL001	Proportion of all applicants of total applicants whose age is known who are aged 20 & under
	APL002	Proportion of all applicants of total applicants whose age is known who are aged 21 - 24
	APL003	Proportion of all applicants of total applicants whose age is known who are aged 25 & above
	APL004	Proportion of successful applicants whose age is known who are aged 20 & under
	APL005	Proportion of successful applicants whose age is known who are aged 21 - 24
	APL006	Proportion of successful applicants whose age is known who are aged 25 & above
	APL007	Proportion of successful applicants to total applicants whose age is known who are aged 20 & under
	APL008	Proportion of successful applicants to total applicants whose age is known who are aged 21 - 24
	APL009	Proportion of successful applicants to total applicants whose age is known who are aged 25 & above
	APL010	Proportion of all applicants whose domicile is known who are UK domiciles
	APL011	Proportion of all applicants whose domicile is known who are "Other-EU" domiciles
	APL012	Proportion of all applicants whose domicile is known who are non-EU domiciles
	APL013	Proportion of successful applicants whose domicile is known who are UK domiciled
	APL014	Proportion of successful applicants whose domicile is known who "Other-EU" domiciled
	APL015	Proportion of successful applicants whose domicile is known who are non-EU domiciles
	APL016	Proportion of successful applicants to applicants whose domicile is known who are UK domiciles
	APL017	Proportion of successful applicants to applicants whose domicile is known who are "Other-EU" domiciles
	APL018	Proportion of successful applicants to applicants whose domicile is known who are non-EU domiciles
	APL019	Proportion of total applicants whose gender is defined who are female
	APL020	Proportion of total applicants whose gender is defined who are male
	APL021	Proportion of total successful applicants whose gender is defined who are female
	APL022	Proportion of total successful applicants whose gender is defined who are male
	APL023	Proportion of successful female applicants to all female applicants (whose gender is known)
	APL024	Proportion of successful male applicants to all male applicants (whose gender is defined)
	APL025	Total number of applicants
	APL026	Total number of successful applicants
	APL027	Proportion of applicants who are successful
Destinations of leavers from HE	DLH001	Proportion of UK domiciled Postgraduate leavers who obtained qualifications through full-time study and entered employment (including those that are working & studying)
	DLH002	Proportion of UK domiciled First degrees leavers who obtained qualifications through full-time study and entered employment (including those that are working & studying)
	DLH003	Proportion of UK domiciled Other UG leavers who obtained qualifications through full-time study and entered employment (including those that are working & studying)
	DLH004	Proportion of UK domiciled Total leavers who obtained qualifications through full-time study and entered employment (including those that are working & studying)
	DLH005	Proportion of UK domiciled Postgraduate leavers who obtained qualifications through part-time study and entered employment (including those that are working & studying)
	DLH006	Proportion of UK domiciled First degrees leavers who obtained qualifications through part-time study and entered employment (including those that are working & studying)
	DLH007	Proportion of UK domiciled Other UG leavers who obtained qualifications through part-time study and entered employment (including those that are working & studying)
	DLH008	Proportion of UK domiciled Total leavers who obtained qualifications through part-time study and entered employment (including those that are working & studying)

Destinations of leavers from HE	DLH009	Proportion of UK domiciled Postgraduate leavers who obtained qualifications through full-time study and were reported as unemployed
	DLH010	Proportion of UK domiciled First degrees leavers who obtained qualifications through full-time study and were reported as unemployed
	DLH011	Proportion of UK domiciled Other UG leavers who obtained qualifications through full-time study and were reported as unemployed
	DLH012	Proportion of UK domiciled Total leavers who obtained qualifications through full-time study and were reported as unemployed
	DLH013	Proportion of UK domiciled Postgraduate leavers who obtained qualifications through part-time study and were reported as unemployed
	DLH014	Proportion of UK domiciled First degrees leavers who obtained qualifications through part-time study and were reported as unemployed
	DLH015	Proportion of UK domiciled Other UG leavers who obtained qualifications through part-time study and were reported as unemployed
	DLH016	Proportion of UK domiciled Total leavers who obtained qualifications through part-time study and were reported as unemployed
	DLH017	Proportion of UK domiciled Postgraduate leavers who obtained qualifications through full-time study and entered further study (including those that are working & studying)
	DLH018	Proportion of UK domiciled First degrees leavers who obtained qualifications through full-time study and entered further study (including those that are working & studying)
	DLH019	Proportion of UK domiciled Other UG leavers who obtained qualifications through full-time study and entered further study (including those that are working & studying)
	DLH020	Proportion of UK domiciled Total leavers who obtained qualifications through full-time study and entered further study (including those that are working & studying)
	DLH021	Proportion of UK domiciled Postgraduate leavers who obtained qualifications through part-time study and entered further study (including those that are working & studying)
	DLH022	Proportion of UK domiciled First degrees leavers who obtained qualifications through part-time study and entered further study (including those that are working & studying)
	DLH023	Proportion of UK domiciled Other UG leavers who obtained qualifications through part-time study and entered further study (including those that are working & studying)
	DLH024	Proportion of UK domiciled Total leavers who obtained qualifications through part-time study and entered further study (including those that are working & studying)
	DLH025	Proportion of UK domiciled Postgraduate leavers who obtained qualifications through full-time study and were reported as having an unknown destination
	DLH026	Proportion of UK domiciled First degrees leavers who obtained qualifications through full-time study and were reported as having an unknown destination
	DLH027	Proportion of UK domiciled Other UG leavers who obtained qualifications through full-time study and were reported as having an unknown destination
	DLH028	Proportion of UK domiciled Total leavers who obtained qualifications through full-time study and were reported as having an unknown destination
	DLH029	Proportion of UK domiciled Postgraduate leavers who obtained qualifications through part-time study and were reported as having an unknown destination
	DLH030	Proportion of UK domiciled First degrees leavers who obtained qualifications through part-time study and were reported as having an unknown destination
	DLH031	Proportion of UK domiciled Other UG leavers who obtained qualifications through part-time study and were reported as having an unknown destination
	DLH032	Proportion of UK domiciled Total leavers who obtained qualifications through part-time study and were reported as having an unknown destination
	DLH033	Proportion of UK domiciled Total leavers who obtained qualifications through full-time study and entered employment (including those that are working & studying) or entered further study (including those that are working & studying)
	DLH034	Proportion of UK domiciled Total leavers who obtained qualifications through part-time study and entered employment (including those that are working & studying) or entered further study (including those that are working & studying)
Research Statistics	RES001	Statistic A: Market share of OSI Research Councils (Actual) research grants & contracts income
	RES002	Statistic A: Market share of OSI Research Councils (Adjusted) research grants & contracts income
	RES003	Statistic A: Market share of UK-based charities (Actual) research grants & contracts income
	RES004	Statistic A: Market share of UK-based charities (Adjusted) research grants & contracts income

Research Statistics	RES005	Statistic A: Market share of UK cent govt/local, health & hospital authorities (Actual) research grants & contracts income
	RES006	Statistic A: Market share of UK cent govt/local, health & hospital authorities (Adjusted) research grants & contracts income
	RES007	Statistic A: Market share of UK industry, commerce & public corporations (Actual) research grants & contracts income
	RES008	Statistic A: Market share of UK industry, commerce & public corporations (Adjusted) research grants & contracts income
	RES009	Statistic A: Market share of EU sources (Actual) research grants & contracts income
	RES010	Statistic A: Market share of EU sources (Adjusted) research grants & contracts income
	RES011	Statistic A: Market share of Other overseas sources (Actual) research grants & contracts income
	RES012	Statistic A: Market share of Other overseas sources (Adjusted) research grants & contracts income
	RES013	Statistic A: Market share of Other sources (Actual) research grants & contracts income
	RES014	Statistic A: Market share of Other sources (Adjusted) research grants & contracts income
	RES015	Statistic A: Total Actual market share of research grants & contracts income
	RES016	Statistic A: Total Adjusted market share of research grants & contracts income
	RES017	Statistic C: Market share of research staff and research council research studentships - Research council research studentships
	RES018	Statistic D: Proportion of total research income from external sources (%)(6)
	RES019	Statistic E: Proportion of total academic income earned for research (%)(6)
	RES020	Funding council recurrent grant for research (£000s)
	RES021	Statistic B: Market share of research staff and research council research studentships - Teaching & research/ research only staff
UCAS Statistics	UCA001	Number of (HESA All) new entrants
	UCA002	Number of (HESA) Full-time first degree/DipHE/HND entrants
	UCA003	Number of UCAS Acceptances
	UCA004	Institutional distribution of applications by domicile UK
	UCA005	Institutional distribution of applications by domicile Other EU
	UCA006	Institutional distribution of applications by domicile Non-EU
	UCA007	Market share of applications by domicile UK
	UCA008	Market share of applications by domicile Other EU
	UCA009	Market share of applications by domicile Non-EU
	UCA010	Market share of applications by domicile Total
	UCA011	Institutional distribution of accepted applicants by domicile UK
	UCA012	Institutional distribution of accepted applicants by domicile Other EU
	UCA013	Institutional distribution of accepted applicants by domicile Non-EU
	UCA014	Market share of accepted applicants by domicile UK
	UCA015	Market share of accepted applicants by domicile Other EU
	UCA016	Market share of accepted applicants by domicile Non-EU
	UCA017	Market share of accepted applicants by domicile Total
	UCA018	Institutional distribution of applications by age 20 years & under
	UCA019	Institutional distribution of applications by age 21 to 24 years
	UCA020	Institutional distribution of applications by age 25 years & above
	UCA021	Market share of applications by age 20 years & under
	UCA022	Market share of applications by age 21 to 24 years
	UCA023	Market share of applications by age 25 years & above
	UCA024	Market share of applications by age Total
	UCA025	Institutional distribution of accepted applicants by age 20 years & under
	UCA026	Institutional distribution of accepted applicants by age 21 to 24 years
	UCA027	Institutional distribution of accepted applicants by age 25 years & above
	UCA028	Market share of accepted applicants by age 20 years & under
	UCA029	Market share of accepted applicants by age 21 to 24 years
	UCA030	Market share of accepted applicants by age 25 years & above
	UCA031	Market share of accepted applicants by age Total
	UCA032	Institutional distribution of applications by gender Male

UCAS Statistics	UCA033	Institutional distribution of applications by gender Female
	UCA034	Market share of applications by gender Male
	UCA035	Market share of applications by gender Female
	UCA036	Market share of applications by gender Total
	UCA037	Institutional distribution of accepted applicants by gender Male
	UCA038	Institutional distribution of accepted applicants by gender Female
	UCA039	Market share of accepted applicants by gender Male
	UCA040	Market share of accepted applicants by gender Female
	UCA041	Market share of accepted applicants by gender Total
Unit Expenditure Statistics	UEX001	Total academic departments: Statistic A: Total expenditure per FTE student
	UEX002	Total academic departments: Statistic B: % spent on academic staff
	UEX003	Total academic departments: Statistic C: % spent on non-academic staff
	UEX004	Total academic departments: Statistic D: % spent on non-staff costs
	UEX005	Statistic F: Total academic services expenditure per FTE student
	UEX006	Administration & central services: Statistic K: Total administration & central services expenditure per FTE student(3)
	UEX007	Administration & central services: Statistic L: % spent on central administration & services
	UEX008	Administration & central services: Statistic M: % spent on general educational expenditure
	UEX009	Administration & central services: Statistic N: % spent on staff & student facilities
	UEX010	Premises: Statistic O: Total premises expenditure per FTE student
	UEX011	Premises: Statistic P: % spent on staff costs
	UEX012	Premises: Statistic Q: % spent on non-staff costs
	UEX013	Premises: Statistic U: % spent on repairs & maintenance
	UEX014	Premises: Statistic W: % spent on other expenditure
HESA Performance Indicators	PIC073	Full-time first degree leavers in the eligible population
	PIC074	Full-time first degree respondents to DLHE survey
	PIC075	Percent of full-time first degree leavers who responded to the DLHE survey
	PIC076	Full-time first degree leavers in the base population
	PIC077	Number of full-time first degree leavers in the base population who were employed, studying or both.
	PIC078	Percent of full-time first degree leavers who were employed, studying or both
	PIC079	Benchmark - full-time first degree leavers who were employed, studying or both
	PIC080	Percent of full-time first degree respondents who were not available for work
	PIC081	Percent of full-time first degree respondents who refused to take part in the survey
	PIC082	Percent of full-time first degree target population who were medical, dental or veterinary graduates
	PIC083	Percent of full-time first degree target population who were on sandwich courses
	PIC084	Number of PhDs awarded per academic staff costs
	PIC085	Research grants & contracts income per academic staff costs
	PIC086	Number of cost centres with academic staff costs
	PIC087	Measure of specialisation per academic staff costs
	PIC088	Number of PhDs awarded per funding council research allocation
	PIC089	Research grants & contracts income per funding council research allocation
	PIC090	Number of cost centres with funding council funding for research
	PIC091	Measure of specialisation per funding council research allocation
	PIC092	QR research funding (£)
	PIC160	Number of PhDs awarded
	PIC093	Number of full-time undergraduate entrants
	PIC094	Number of young full-time undergraduate entrants
	PIC095	Percent of full-time undergraduate entrants who are young
	PIC096	Number of young full-time undergraduate entrants with known state school data
	PIC097	Percent of young full-time undergraduate entrants with known state school data
	PIC098	Number of young full-time undergraduate entrants from state schools or colleges
	PIC099	Indicator - Percent of young full-time undergraduate entrants from state schools or colleges
	PIC100	Benchmark - young full-time undergraduate entrants from state schools or colleges

HESA Performance Indicators	PIC101	Location adjusted benchmark - young full-time undergraduate entrants from state schools or colleges
	PIC102	Number of young full-time undergraduate entrants with known NS-SEC data
	PIC103	Percent of young full-time undergraduate entrants with known NS-SEC data
	PIC104	Number of young full-time undergraduate entrants from NS-SEC classes 4,5,6 & 7
	PIC105	Percent of young full-time undergraduate entrants from NS-SEC classes 4,5,6 & 7
	PIC106	Benchmark - young full-time undergraduate entrants from NS-SEC classes 4,5,6 & 7
	PIC107	Location adjusted benchmark - young full-time undergraduate entrants from NS-SEC classes 4,5,6 & 7
	PIC108	Number of young full-time undergraduate entrants with known participation neighbourhood data
	PIC109	Percent of young full-time undergraduate entrants with known participation neighbourhood data
	PIC110	Number of young full-time undergraduate entrants from low participation neighbourhoods
	PIC111	Percent of young full-time undergraduate entrants from low participation neighbourhoods
	PIC112	Benchmark - young full-time undergraduate entrants from low participation neighbourhoods
	PIC113	Location adjusted benchmark - young full-time undergraduate entrants from low participation neighbourhoods
	PIC114	Number of full-time first degree entrants
	PIC115	Number of mature full-time first degree entrants
	PIC116	Percent of full-time first degree entrants who are mature
	PIC117	Number of mature full-time first degree entrants with known highest qualification on entry & participation neighbourhood data
	PIC118	Percent of mature full-time first degree entrants with known highest qualification on entry & participation neighbourhood data
	PIC119	Number of mature full-time first degree entrants whose highest qualification on entry isn't HE
	PIC120	Number of mature full-time first degree entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC121	Percent of mature full-time first degree entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC122	Benchmark - mature full-time first degree entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC123	Location adjusted benchmark - mature full-time first degree entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC124	Number of full-time undergraduate entrants
	PIC125	Number of mature full-time undergraduate entrants
	PIC126	Percent of full-time undergraduate entrants who are mature
	PIC127	Number of mature full-time undergraduate entrants with known highest qualification on entry & participation neighbourhood data
	PIC128	Percent of mature full-time undergraduate entrants with known highest qualification on entry & participation neighbourhood data
	PIC129	Number of mature full-time undergraduate entrants whose highest qualification on entry isn't HE
	PIC130	Number of mature full-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC131	Percent of mature full-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC132	Benchmark - mature full-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC133	Location adjusted benchmark - mature full-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC134	Number of young part-time undergraduate entrants
	PIC135	Percent of part-time undergraduate entrants who are young
	PIC136	Number of young part-time undergraduate entrants with known highest qualification on entry & participation neighbourhood data
	PIC137	Percent of young undergraduate entrants with known highest qualification on entry & participation neighbourhood data

HESA Performance Indicators	PIC138	Number of young part-time undergraduate entrants whose highest qualification on entry isn't HE
	PIC139	Number of young part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC140	Percent of young part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC141	Benchmark - young part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC142	Location adjusted benchmark - young part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC143	Number of mature part-time undergraduate entrants
	PIC144	Percent of part-time undergraduate entrants who are mature
	PIC145	Number of mature part-time undergraduate entrants with known highest qualification on entry & participation neighbourhood data
	PIC146	Percent of mature undergraduate entrants with known highest qualification on entry & participation neighbourhood data
	PIC147	Number of mature part-time undergraduate entrants whose highest qualification on entry isn't HE
	PIC148	Number of mature part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC149	Percent of mature part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC150	Benchmark - mature part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC151	Location adjusted benchmark - mature part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC152	Number of part-time undergraduate entrants
	PIC153	Number of part-time undergraduate entrants with known highest qualification on entry & participation neighbourhood data
	PIC154	Percent of undergraduate entrants with known highest qualification on entry & participation neighbourhood data
	PIC155	Number of part-time undergraduate entrants whose highest qualification on entry isn't HE
	PIC156	Number of part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC157	Percent of part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC158	Benchmark - Part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC159	Location adjusted benchmark - Part-time undergraduate entrants from low participation neighbourhoods whose highest qualification on entry isn't HE
	PIC001	Number of young full-time undergraduate entrants last year
	PIC002	Number of young full-time undergraduate entrants last year who continue or qualify at same HEI
	PIC003	Percent of young full-time undergraduate entrants last year who continue or qualify at same HEI
	PIC004	Benchmark - young full-time first degree entrants who continue or qualify at same HEI
	PIC005	Benchmark - young full-time other undergraduate entrants who continue or qualify at same HEI
	PIC006	Number of young full-time undergraduate entrants who transfer to other UK HEI
	PIC007	Percent of young full-time first degree entrants who transfer to other UK HEI
	PIC008	Benchmark - young full-time first degree entrants who transfer to other UK HEI
	PIC009	Benchmark - young full-time other undergraduate entrants who transfer to other UK HEI
	PIC010	Number of young full-time undergraduate entrants who are no longer in HE
	PIC011	Percent of young full-time undergraduate entrants who are no longer in HE
	PIC012	Benchmark - young full-time first degree entrants who are no longer in HE
	PIC013	Benchmark - young full-time other undergraduate entrants who are no longer in HE
	PIC014	Number of mature full-time undergraduate entrants
	PIC015	Number of mature full-time undergraduate entrants who continue or qualify at same HEI
	PIC016	Percent of mature full-time first degree entrants who continue or qualify at same HEI
	PIC017	Benchmark - mature full-time first degree entrants who continue or qualify at same HEI

HESA Performance Indicators	PIC018	Benchmark - mature full-time other undergraduate entrants who continue or qualify at same HEI
	PIC019	Number of mature full-time undergraduate degree entrants who transfer to other UK HEI
	PIC020	Percentage of mature full-time undergraduate degree entrants who transfer to other UK HEI
	PIC021	Benchmark - mature full-time first degree entrants who transfer to other UK HEI
	PIC022	Benchmark - mature full-time other undergraduate entrants who transfer to other UK HEI
	PIC023	Number of mature full-time undergraduate entrants who are no longer in HE
	PIC024	Percent of mature full-time undergraduate entrants who continue or qualify at same HEI
	PIC025	Benchmark - mature full-time first degree entrants who are no longer in HE
	PIC026	Benchmark - mature full-time other undergraduate entrants who are no longer in HE
	PIC027	Number of full-time undergraduate entrants
	PIC028	Number of full-time undergraduate entrants who continue or qualify at same HEI
	PIC029	Percent of full-time undergraduate entrants who continue or qualify at same HEI
	PIC030	Benchmark - full-time first degree entrants who continue or qualify at same HEI
	PIC031	Benchmark - full-time other undergraduate entrants who continue or qualify at same HEI
	PIC032	Number of full-time undergraduate entrants who transfer to other UK HEI
	PIC033	Percent of full-time undergraduate entrants who transfer to other UK HEI
	PIC034	Benchmark - full-time first degree entrants who transfer to other UK HEI
	PIC035	Benchmark - full-time other undergraduate entrants who transfer to other UK HEI
	PIC036	Number of full-time undergraduate entrants who are no longer in HE
	PIC037	Percent of full-time undergraduate entrants who are no longer in HE
	PIC038	Benchmark - full-time first degree entrants who are no longer in HE
	PIC039	Benchmark - full-time other undergraduate entrants who are no longer in HE
	PIC040	Number of young undergraduate degree entrants not in HE the year after they entered
	PIC041	Number of young undergraduate degree entrants not in HE the year after they entered & who resume at the same HEI
	PIC042	Percent of young undergraduate degree entrants not in HE the year after they entered & who resume at the same HEI
	PIC043	Number of young undergraduate degree entrants not in HE the year after they entered & who transfer to another UK HEI
	PIC044	Percent of young undergraduate degree entrants not in HE the year after they entered & who transfer to another UK HEI
	PIC045	Number of young undergraduate degree entrants not in HE for two consecutive years
	PIC046	Percent of young undergraduate degree entrants not in HE for two consecutive years
	PIC047	Number of mature undergraduate degree entrants not in HE the year after they entered
	PIC048	Number of mature undergraduate degree entrants not in HE the year after they entered & who resume at the same HEI
	PIC049	Percent of mature undergraduate degree entrants not in HE the year after they entered & who resume at the same HEI
	PIC050	Number of mature undergraduate degree entrants not in HE the year after they entered & who transfer to another UK HEI
	PIC051	Percent of mature undergraduate degree entrants not in HE the year after they entered & who transfer to another UK HEI
	PIC052	Number of mature undergraduate degree entrants not in HE for two consecutive years
	PIC053	Percent of mature undergraduate degree entrants not in HE for two consecutive years
	PIC054	Number of undergraduate degree entrants not in HE the year after they entered
	PIC055	Number of undergraduate degree entrants not in HE the year after they entered & who resume at the same HEI
	PIC056	Percent of undergraduate degree entrants not in HE the year after they entered & who resume at the same HEI
	PIC057	Number of undergraduate degree entrants not in HE the year after they entered & who transfer to another UK HEI
	PIC058	Percent of undergraduate degree entrants not in HE the year after they entered & who transfer to another UK HEI
	PIC059	Number of undergraduate degree entrants not in HE for two consecutive years
	PIC060	Percent of undergraduate degree entrants not in HE for two consecutive years
	PIC061	Number of full-time first degree students

HESA Performance Indicators	PIC062	Number of full-time first degree students who are in receipt of DSA
	PIC063	Percent of full-time first degree students who are in receipt of DSA
	PIC064	Benchmark - full-time first degree students who are in receipt of DSA
	PIC065	Number of full-time undergraduate students
	PIC066	Number of full-time undergraduate students who are in receipt of DSA
	PIC067	Percent of full-time undergraduate students who are in receipt of DSA
	PIC068	Benchmark - full-time undergraduate students who are in receipt of DSA
	PIC069	Number of part-time undergraduate students
	PIC070	Number of part-time undergraduate students who are in receipt of DSA
	PIC071	Percent of part-time undergraduate students who are in receipt of DSA
	PIC072	Benchmark - part-time undergraduate students who are in receipt of DSA
Student-Staff Ratios	SSR001	Ratio by Institution
HESA Finance	UFI001	Consolidated income and expenditure account - Expenditure - Exceptional items
	UFI002	Consolidated income and expenditure account - Expenditure - Staff costs
	UFI003	Consolidated income and expenditure account - Expenditure - Other operating expenses
	UFI004	Consolidated income and expenditure account - Expenditure - Depreciation
	UFI005	Consolidated income and expenditure account - Expenditure - Interest & other finance costs
	UFI006	Consolidated income and expenditure account - Expenditure - Total Expenditure
	UFI007	Consolidated income and expenditure account - Income - Funding body grants
	UFI008	Consolidated income and expenditure account - Income - Tuition fees & education contracts
	UFI009	Consolidated income and expenditure account - Income - Research grants & contracts
	UFI010	Consolidated income and expenditure account - Income - Other income
	UFI011	Consolidated income and expenditure account - Income - Endowment & investment income
	UFI012	Consolidated income and expenditure account - Income - Total Income (group & share of joint venture(s))
	UFI013	Consolidated income and expenditure account - Minority interest - Total
	UFI014	Consolidated income and expenditure account - Note of group historical cost surpluses & deficits for the year ended 31 July - Surplus/(deficit) on continuing operations before taxation
	UFI015	Consolidated income and expenditure account - Note of group historical cost surpluses & deficits for the year ended 31 July - Difference between historical cost depreciation & the actual charge for the year calculated on the re-valued amount
	UFI016	Consolidated income and expenditure account - Note of group historical cost surpluses & deficits for the year ended 31 July - Realisation of property revaluation gains of previous years
	UFI017	Consolidated income and expenditure account - Note of group historical cost surpluses & deficits for the year ended 31 July - Historical cost surplus/(deficit) for the year before taxation
	UFI018	Consolidated income and expenditure account - Note of group historical cost surpluses & deficits for the year ended 31 July - Historical cost surplus/(deficit) for the year after taxation
	UFI019	Consolidated income and expenditure account - Surplus/(deficit) for the year retained within general reserves - Total
	UFI020	Consolidated income and expenditure account - Taxation - Total
	UFI021	Consolidated income and expenditure account - Transfer from/(to) accumulated income in endowment funds - Total
	UFI022	Balance sheet - Creditors: Amounts falling due after more one year - Reimbursable to the Funding Council Balance as at 31 July
	UFI023	Balance sheet - Creditors: Amounts falling due after more one year - External borrowing Balance as at 31 July
	UFI024	Balance sheet - Creditors: Amounts falling due after more one year - Other Balance as at 31 July
	UFI025	Balance sheet - Creditors: Amounts falling due after more one year - Total Creditors Balance as at 31 July

HESA Finance	UFI026	Balance sheet - Creditors: Amounts falling due within one year - Creditors Balance as at 31 July current year
	UFI027	Balance sheet - Creditors: Amounts falling due within one year - Current portion of long-term liabilities Balance as at 31 July
	UFI028	Balance sheet - Creditors: Amounts falling due within one year - Bank overdrafts Balance as at 31 July
	UFI029	Balance sheet - Creditors: Amounts falling due within one year - Total Creditors Balance as at 31 July
	UFI030	Balance sheet - Current assets - Stock Balance as at 31 July
	UFI031	Balance sheet - Current assets - Debtors Balance as at 31 July
	UFI032	Balance sheet - Current assets - Investments Balance as at 31 July
	UFI033	Balance sheet - Current assets - Cash at bank & in hand Balance as at 31 July
	UFI034	Balance sheet - Current assets - Total Current Assets Balance as at 31 July
	UFI035	Balance sheet - Deferred capital grants - Balance as at 31 July
	UFI036	Balance sheet - Endowment Assets - Balance as at 31 July
	UFI037	Balance sheet - Endowments - Total Endowments Balance as at 31 July
	UFI038	Balance sheet - Fixed Assets - Tangible assets Balance as at 31 July
	UFI039	Balance sheet - Fixed Assets - Total Fixed Assets Balance as at 31 July
	UFI040	Balance sheet - Net assets including pension asset/(liability) - Balance as at 31 July
	UFI041	Balance sheet - Net current assets/(liabilities) - Balance as at 31 July
	UFI042	Balance sheet - Provisions for liabilities and charges - Balance as at 31 July
	UFI043	Balance sheet - Reserves - Income & expenditure account Balance as at 31 July
	UFI044	Balance sheet - Total assets less current liabilities - Balance as at 31 July
	UFI045	Balance sheet - Total funds - Balance as at 31 July
	UFI046	Cash flow statement - Capital expenditure and financial investment - Payments to acquire tangible assets Year ended 31 July
	UFI047	Cash flow statement - Capital expenditure and financial investment - Payments to acquire endowment asset investments Year ended 31 July
	UFI048	Cash flow statement - Capital expenditure and financial investment - Total payments to acquire fixed/endowment assets Year ended 31 July
	UFI049	Cash flow statement - Capital expenditure and financial investment - Receipts from sale of tangible assets Year ended 31 July
	UFI050	Cash flow statement - Capital expenditure and financial investment - Receipts from sale of endowment assets Year ended 31 July
	UFI051	Cash flow statement - Capital expenditure and financial investment - Deferred capital grants received Year ended 31 July
	UFI052	Cash flow statement - Capital expenditure and financial investment - Endowments received Year ended 31 July
	UFI053	Cash flow statement - Capital expenditure and financial investment - Other items Year ended 31 July
	UFI054	Cash flow statement - Capital expenditure and financial investment - Net cash inflow/(outflow) from capital expenditure & financial investment Year ended 31 July
	UFI055	Cash flow statement - Capital expenditure and financial investment - Capital element of finance lease repayments Year ended 31 July
	UFI056	Cash flow statement - Financing - Mortgages & loans acquired Year ended 31 July
	UFI057	Cash flow statement - Financing - Mortgage & loan capital repayments Year ended 31 July
	UFI058	Cash flow statement - Financing - Other items Year ended 31 July
	UFI059	Cash flow statement - Financing - Net cash inflow/(outflow) from financing Year ended 31 July
	UFI060	Cash flow statement - Increase/(decrease) in cash in the year - Year ended 31 July
	UFI061	Cash flow statement - Management of liquid resources - Year ended 31 July
	UFI062	Cash flow statement - Reconciliation of net cash flow movement in net funds/(debt) - Increase/(decrease) in cash in the year ended 31 July
	UFI063	Cash flow statement - Reconciliation of net cash flow movement in net funds/(debt) - Net funds/(debt) at 1 August Year ended 31 July
	UFI064	Cash flow statement - Returns on investments and servicing of finance - Income from endowments Year ended 31 July
	UFI065	Cash flow statement - Returns on investments and servicing of finance - Income from short-term investments Year ended 31 July

Finance	UFI066	Cash flow statement - Returns on investments and servicing of finance - Other interest received Year ended 31 July
	UFI067	Cash flow statement - Returns on investments and servicing of finance - Interest paid Year ended 31 July
	UFI068	Cash flow statement - Returns on investments and servicing of finance - Other items Year ended 31 July
	UFI069	Cash flow statement - Returns on investments and servicing of finance - Net cash inflow/(outflow) from returns on investments & servicing of finance Year ended 31 July
	UFI070	Cash flow statement - Taxation - Total
	UFI071	Tuition fees & education contracts analysed by domicile, mode, level & source - FE Course Fees - Total
	UFI072	Tuition fees & education contracts analysed by domicile, mode, level & source - Non-Credit bearing Course Fees - Total
	UFI073	Tuition fees & education contracts analysed by domicile, mode, level & source - Research Training Support Grants - Total research training support grants
	UFI074	Tuition fees & education contracts analysed by domicile, mode, level & source - Total Tuition Fees and Education Contracts - Total
	UFI075	Income analysed by source - Other Income - Residences & catering operations (including conferences)
	UFI076	Income analysed by source - Other Income - Grants from local authorities
	UFI077	Income analysed by source - Other Income - Income from health & hospital authorities (excluding teaching contracts for student provision)
	UFI078	Income analysed by source - Other Income - Release of deferred capital grants
	UFI079	Income analysed by source - Other Income - Income from intellectual property rights
	UFI080	Income analysed by source - Other Income - Other operating income
	UFI081	Income analysed by source - Other Income - Total Other Income
	UFI082	Income analysed by source - Total Income
	UFI083	Income analysed by source - Total Tuition Fees and Education Contracts - Total
	UFI084	Capital expenditure - Total Capital Expenditure - Total Actual Spend
	UFI085	Capital expenditure - Total Capital Expenditure - Funding Council grants
	UFI086	Capital expenditure - Total Capital Expenditure - Retained proceeds of sales
	UFI087	Capital expenditure - Total Capital Expenditure - Internal funds
	UFI088	Capital expenditure - Total Capital Expenditure - Loans
	UFI089	Capital expenditure - Total Capital Expenditure - Other external sources
HESA - KFIs	KFI001	1 - Total income in £000s
	KFI002	2 - Percentage ratio of total funding body grants to total income
	KFI003	3 - Percentage ratio of recurrent teaching grants from funding bodies for HE provision to total income
	KFI004	4 - Percentage ratio of recurrent research grants from funding bodies for HE provision to total income
	KFI005	5 - Percentage ratio of recurrent (other) grants from funding bodies for HE provision to total income
	KFI006	6 - Percentage ratio of total funding body grants for HE provision to total income
	KFI008	8 - Percentage ratio of tuition fees & education contracts to total income
	KFI009	9 - Percentage ratio of full-time home/European Union (EU) HE student fees to total income
	KFI010	10 - Percentage ratio of non-EU HE course fees to total income
	KFI011	11 - Percentage ratio of income from research grants & contracts to total income
	KFI012	12 - Percentage ratio of Department for Business, Innovation & Skills (BIS) research council grants & contracts to total income
	KFI014	14 - Percentage ratio of UK industry, commerce & public corporations research grants & contracts to total income
	KFI016	16 - Percentage ratio of EU research grants & contracts to total income
	KFI018	18 - Percentage ratio of other income to total income
	KFI019	19 - Percentage ratio of total endowment & investment income to total income
	KFI020	20 - Percentage ratio of contribution from research grants & contracts to research grants & contracts income
	KFI022	22a - Percentage ratio of historical surplus/(deficit) for the year after taxation to total income
	KFI024	23 - Ratio of current assets to current liabilities

HESA - KFIs	KFI025	24 - Ratio of liquid assets to current liabilities
	KFI027	26a - Days ratio of net liquidity to total expenditure (excluding depreciation)
	KFI029	27a - Days ratio of total general funds to total expenditure
	KFI031	28 - Days of current income (excluding funding body grants for HE provision (SFC for all provision)) represented by debtors
	KFI032	29 - Percentage ratio of total long-term borrowings to total income
	KFI033	30a - Percentage ratio of total staff costs to total income
	KFI035	31 - Percentage ratio of total net cash inflow from operating activities to total income
	KFI036	32a - Gearing ratio
	KFI038	33a - Percentage ratio of premises repairs & maintenance to total expenditure
	KFI040	34a - Days ratio of total net cash inflow to total expenditure (excluding depreciation)
	KFI043	36 - Percentage ratio of interest & other finance costs to total income
Staffing	STA001	Number of academic staff (FPE) leavers (exc. atypical) who are Teaching only
	STA002	Number of academic staff (FPE) leavers (exc. atypical) who are Research only
	STA003	Number of academic staff (FPE) leavers (exc. atypical) who are Teaching & research
	STA004	Number of academic staff (FPE) leavers (exc. atypical) who are Neither teaching nor research
	STA042	Proportion of academic staff (FPE) leavers (exc. atypical) who are Teaching only
	STA043	Proportion of academic staff (FPE) leavers (exc. atypical) who are Research only
	STA044	Proportion of academic staff (FPE) leavers (exc. atypical) who are Teaching & research
	STA005	Number of academic staff (FPE) leavers (exc. atypical) who are Full-time
	STA006	Number of academic staff (FPE) leavers (exc. atypical) who are Part-time
	STA007	Proportion of all academic leavers (exc. atypical) who were in full-time employment
	STA008	Proportion of all academic leavers (exc. atypical) who were in part-time employment
	STA009	Number of academic staff (FPE) starters (exc. atypical) who are Teaching only
	STA010	Number of academic staff (FPE) starters (exc. atypical) who are Research only
	STA011	Number of academic staff (FPE) starters (exc. atypical) who are Teaching & research
	STA012	Number of academic staff (FPE) starters (exc. atypical) who are Neither teaching nor research
	STA045	Proportion of academic staff (FPE) starters (exc. atypical) who are Teaching only
	STA046	Proportion of academic staff (FPE) starters (exc. atypical) who are Research only
	STA047	Proportion of academic staff (FPE) starters (exc. atypical) who are Teaching & research
	STA013	Number of academic staff (FPE) starters (exc. atypical)who are Full-time
	STA014	Number of academic staff (FPE) starters (exc. atypical) who are Part-time
	STA015	Proportion of all academic starters (exc. atypical) who were in full-time employment
	STA016	Proportion of all academic starters (exc. atypical) who were in part-time employment
	STA055	Growth in number of academic staff (FPE - exc. atypical) who are Teaching only
	STA056	Growth in number of academic staff (FPE - exc. atypical) who are Research only
	STA057	Growth in number of academic staff (FPE - exc. atypical) who are Teaching & research
	STA058	Growth in number of academic staff (FPE - exc. atypical) who are Neither teaching nor research
	STA059	Growth in number of academic staff (FPE - exc. atypical) who are Full-time
	STA060	Growth in number of academic staff (FPE - exc. atypical) who are Part-time
	STA017	Number of staff (FTE) who are Teaching only
	STA018	Number of staff (FTE) who are Research only
	STA019	Number of staff (FTE) who are Teaching & research
	STA020	Number of staff (FTE) who are Neither teaching nor research
	STA021	Number of staff (FTE) who are Not applicable/Not required (Default code)
	STA048	Proportion of staff (FTE) who are Teaching only
	STA049	Proportion of staff (FTE) who are Research only
	STA050	Proportion of staff (FTE) who are Teaching & research
	STA051	Proportion of staff (FTE) who are Neither teaching nor research
	STA022	Number of staff (FTE) who are Non academic
	STA023	Number of staff (FTE) who are Academic
	STA052	Proportion of staff (FTE) who are Academic
	STA024	Number of staff (FTE) who are full-time

Staffing	STA025	Number of staff (FTE) who are part-time
	STA026	Proportion of staff who are full-time
	STA027	Number of staff (FTE) who are of UK nationality
	STA028	Number of staff (FTE) who are of "Other EU" nationality
	STA029	Number of staff (FTE) who are of Non-EU nationality
	STA030	Number of staff (FTE) whose nationality is unknown
	STA053	Proportion of staff (FTE) whose nationality is known who are of UK nationality
	STA054	Proportion of staff (FTE) whose nationality is known who are of "Other-EU" nationality
	STA031	Number of staff (FTE) who are Wholly institutionally financed
	STA032	Number of staff (FTE) who are Principally financed by the institution
	STA033	Number of staff (FTE) who are Other sources of finance
	STA034	Number of staff (FTE) who are Not applicable/not required
	STA061	Proportion of staff (FTE) who are Wholly institutionally financed
	STA062	Proportion of staff (FTE) who are Principally financed by the institution
	STA063	Proportion of staff (FTE) who are Other sources of finance
	STA035	Number of staff (FTE) who are Open-ended/permanent
	STA036	Number of staff (FTE) who are Fixed-term contract
	STA037	Number of staff (FTE) who are Atypical
	STA064	Proportion of staff (FTE) who are Open-ended/permanent
	STA065	Proportion of staff (FTE) who are Fixed-term contract
	STA038	Number of staff (FTE) who are typical
	STA039	Number of staff (FTE) who are Atypical
	STA066	Proportion of staff (FTE) who are typical
Students	STU001	Total FTE Student Instance Count
	STU005	Number of HE student qualifiers (FPE) who are 17 and under
	STU006	Number of HE student qualifiers (FPE) who are 18-20 years
	STU007	Number of HE student qualifiers (FPE) who are 21-24 years
	STU008	Number of HE student qualifiers (FPE) who are 25-29 years
	STU009	Number of HE student qualifiers (FPE) who are 30 and over
	STU010	Number of HE student qualifiers (FPE) who are Age unknown
	STU011	Proportion of all students whose age is known who are aged 17 and under
	STU012	Proportion of all students whose age is known who are aged 18 - 20 years
	STU013	Proportion of all students whose age is known who are aged 21 - 24 years
	STU014	Proportion of all students whose age is known who are aged 25 - 29 years
	STU016	Number of HE student qualifiers (FPE) who are First class honours
	STU017	Number of HE student qualifiers (FPE) who are Upper second class honours
	STU018	Number of HE student qualifiers (FPE) who are Lower second class honours
	STU019	Number of HE student qualifiers (FPE) who are Third class honours / Pass
	STU020	Number of HE student qualifiers (FPE) who are Unclassified
	STU021	Number of HE student qualifiers (FPE) who are Classification not applicable
	STU022	Proportion of all FPE HE Student Qualifiers who achieve a First Class Honours degree (out of all applicable degrees)
	STU023	Proportion of all FPE HE Student Qualifiers who achieve an Upper Second Class Honours degree (out of all applicable degrees)
	STU024	Proportion of all FPE HE Student Qualifiers who achieve a Lower Second Class Honours degree (out of all applicable degrees)
	STU025	Proportion of all FPE HE Student Qualifiers who achieve a Third Class Honours degree (out of all applicable degrees)
	STU027	Number of HE student qualifiers (FPE) who are Known to be disabled
	STU028	Number of HE student qualifiers (FPE) who are No known disability
	STU029	Number of HE student qualifiers (FPE) who are Not known/sought
	STU030	Proportion with disability status known who are disabled
	STU031	Number of HE student qualifiers (FPE) who are United Kingdom
	STU032	Number of HE student qualifiers (FPE) who are Other European Union
	STU033	Number of HE student qualifiers (FPE) who are Non-European-Union
	STU034	Proportion of HE Student Qualifiers (FPE) who are domiciled in the UK

Students	STU035	Proportion of HE Student Qualifiers (FPE) who are domiciled in the EU
	STU037	Number of HE student qualifiers (FPE) who are White
	STU038	Number of HE student qualifiers (FPE) who are Black and Minority Ethnic
	STU039	Number of HE student qualifiers (FPE) who are Not known
	STU040	Number of HE student qualifiers (FPE) who are Non-UK domicile
	STU041	Proportion of HE student qualifiers (FPE) whose ethnicity known and is White (UK)
	STU042	Proportion of HE student qualifiers (FPE) whose ethnicity known and is BME (UK)
	STU044	Number of HE student qualifiers (FPE) who are female
	STU045	Number of HE student qualifiers (FPE) who are male
	STU046	Proportion of HE student qualifiers (FPE) whose sex is determinate and who are male
	STU048	Number of HE student qualifiers (FPE) who are UK
	STU049	Number of HE student qualifiers (FPE) who are Other EU
	STU050	Number of HE student qualifiers (FPE) who are Non-EU
	STU051	Number of HE student qualifiers (FPE) who are Unknown
	STU052	Proportion of HE student qualifiers (FPE) whose nationality is known and who are from the UK
	STU053	Proportion of HE student qualifiers (FPE) whose nationality is known and who are from the Other EU
	STU055	Number of HE student qualifiers (FPE) who are Full-time
	STU056	Number of HE student qualifiers (FPE) who are Part-time
	STU057	Proportion of HE student qualifiers (FPE) who are full-time
	STU059	Number of HE students (FPE) who are 18 and under
	STU060	Number of HE students (FPE) who are 19 years
	STU061	Number of HE students (FPE) who are 20 years
	STU062	Number of HE students (FPE) who are 21 - 24
	STU063	Number of HE students (FPE) who are 25 - 29
	STU064	Number of HE students (FPE) who are 30 and over
	STU065	Number of HE students (FPE) who are Age unknown
	STU066	Proportion of HE students whose age is known who are aged 18 and under
	STU067	Proportion of HE students whose age is known who are aged 19 years
	STU068	Proportion of HE students whose age is known who are aged 20 years
	STU069	Proportion of HE students whose age is known who are aged 21 - 24 years
	STU070	Proportion of HE students whose age is known who are aged 25 - 29 years
	STU072	Number of HE students (FPE) who are Known to be disabled
	STU073	Number of HE students (FPE) who are No known disability
	STU074	Number of HE students (FPE) who are Not known/sought
	STU075	Proportion of HE students (FPE) who disability status is known and who are disabled
	STU076	Number of HE students (FPE) who are UK domiciles
	STU077	Number of HE students (FPE) who are Other European Union domiciles
	STU078	Number of HE students (FPE) who are Non-European-Union domiciles
	STU079	Proportion of HE students (FPE) who are UK domiciles
	STU080	Proportion of HE students (FPE) who are Other European Union domiciles
	STU082	Number of HE students (FPE) who are White
	STU083	Number of HE students (FPE) who are Black and Minority Ethnic
	STU084	Number of HE students (FPE) who are Not known
	STU085	Number of HE students (FPE) who are Non-UK domicile
	STU086	Proportion of HE students (FPE) whose ethnicity is known and who are White
	STU087	Proportion of HE students (FPE) whose ethnicity is known and who are Black and Minority Ethnic
	STU089	Number of HE students (FPE) who are Eligible to pay home fees
	STU090	Number of HE students (FPE) who are Not eligible to pay home fees
	STU091	Number of HE students (FPE) who are Eligibility to pay home fees not assessed
	STU092	Number of HE students (FPE) who are First year students
	STU093	Number of HE students (FPE) who are Non-first year students
	STU094	Number of HE student (FPE) who are Female
	STU095	Number of HE student (FPE) who are Male

Students	STU096	Proportion of HE student (FPE) whose sex is determined and who are Male
	STU097	Number of HE students (FPE) who are Postgraduate (research) students
	STU098	Number of HE students (FPE) who are Postgraduate (taught) students
	STU099	Number of HE students (FPE) who are First degree students
	STU100	Number of HE students (FPE) who are Other undergraduate students
	STU101	Proportion of HE students (FPE) who are Postgraduate (research) students
	STU102	Proportion of HE students (FPE) who are Postgraduate (taught) students
	STU103	Proportion of HE students (FPE) who are First degree students
	STU108	Number of HE students (FPE) who are "No award or financial backing" funded
	STU109	Number of HE students (FPE) who are "UK LEA mandatory/discretionary awards" funded
	STU110	Number of HE students (FPE) who are "Institutionally waived/award" funded
	STU111	Number of HE students (FPE) who are "Research councils and British Academy" funded
	STU112	Number of HE students (FPE) who are "Charities and international agencies" funded
	STU113	Number of HE students (FPE) who are "UK central govt/local, health, employment and agriculture authorities/bodies" funded
	STU114	Number of HE students (FPE) who are "EU sources" funded
	STU115	Number of HE students (FPE) who are "Other overseas sources" funded
	STU116	Number of HE students (FPE) who are "UK industry/commerce and students employer" funded
	STU117	Number of HE students (FPE) who are "Absent/no fees" funded
	STU118	Number of HE students (FPE) who are "Not known/Other" funded
	STU119	Proportion of HE students (FPE) whose funding method is known and who are "No award or financial backing" funded
	STU120	Proportion of HE students (FPE) whose funding method is known and who are "UK LEA mandatory/discretionary awards" funded
	STU121	Proportion of HE students (FPE) whose funding method is known and who are "Institutionally waived/award" funded
	STU122	Proportion of HE students (FPE) whose funding method is known and who are "Research councils and British Academy" funded
	STU123	Proportion of HE students (FPE) whose funding method is known and who are "Charities and international agencies" funded
	STU124	Proportion of HE students (FPE) whose funding method is known and who are "UK central govt/local, health, employment and agriculture authorities/bodies" funded
	STU125	Proportion of HE students (FPE) whose funding method is known and who are "EU sources" funded
	STU126	Proportion of HE students (FPE) whose funding method is known and who are "Other overseas sources" funded
	STU127	Proportion of HE students (FPE) whose funding method is known and who are "UK industry/commerce and students employer" funded
	STU129	Number of HE students (FPE) who are Full-time
	STU130	Number of HE students (FPE) who are Part-time
	STU131	Proportion of HE students (FPE) who are Full-time
	STU132	Number of HE students (FPE) who are UK nationals
	STU133	Number of HE students (FPE) who are "Other-EU" nationals
	STU134	Number of HE students (FPE) who are Non-EU nationals
	STU135	Number of HE students (FPE) whose nationality is unknown
	STU136	Proportion of HE students (FPE) whose nationality is known who are UK nationals
	STU137	Proportion of HE students (FPE) whose nationality is known who are "Other-EU" nationals
	STU139	Number of HE students (FPE) who are in their Foundation year
	STU140	Number of HE students (FPE) who are Year 1
	STU141	Number of HE students (FPE) who are Year 2
	STU142	Number of HE students (FPE) who are Year 3
	STU143	Number of HE students (FPE) who are Year 4
	STU144	Number of HE students (FPE) who are Year 5
	STU145	Number of HE students (FPE) who are Year 6+
	STU146	Number of HE students (FPE) whose year of study is unknown
	STU147	Proportion of HE students (FPE) whose year of study is known and who are in their Foundation year

Students	STU148	Proportion of HE students (FPE) whose year of study is known and who are Year 1
	STU149	Proportion of HE students (FPE) whose year of study is known and who are Year 2
	STU150	Proportion of HE students (FPE) whose year of study is known and who are Year 3
	STU151	Proportion of HE students (FPE) whose year of study is known and who are Year 4
	STU152	Proportion of HE students (FPE) whose year of study is known and who are Year 5
	STU154	Number of HE students (FTE) who are aged 18 and under
	STU155	Number of HE students (FTE) who are aged 19 years
	STU156	Number of HE students (FTE) who are aged 20 years
	STU157	Number of HE students (FTE) who are aged 21 - 24
	STU158	Number of HE students (FTE) who are aged 25 - 29
	STU159	Number of HE students (FTE) who are aged 30 and over
	STU160	Number of HE students (FTE) whose age is unknown
	STU161	Proportion of HE students (FTE) whose age is known and who are aged 18 and under
	STU162	Proportion of HE students (FTE) whose age is known and who are aged 19 years
	STU163	Proportion of HE students (FTE) whose age is known and who are aged 20 years
	STU164	Proportion of HE students (FTE) whose age is known and who are aged 21 - 24
	STU165	Proportion of HE students (FTE) whose age is known and who are aged 25 - 29
	STU167	Number of HE students (FTE) who are Known to be disabled
	STU168	Number of HE students (FTE) who are No known disability
	STU169	Number of HE students (FTE) who are Not known/sought
	STU170	Proportion of HE students (FTE) whose disability status is known and who are disabled
	STU171	Number of HE students (FTE) who are UK domiciles
	STU172	Number of HE students (FTE) who are "Other European Union" domiciles
	STU173	Number of HE students (FTE) who are Non-European-Union domiciles
	STU174	Proportion of HE students (FTE) who are UK domiciles
	STU175	Proportion of HE students (FTE) who are "Other European Union" domiciles
	STU177	Number of HE students (FTE) who are White
	STU178	Number of HE students (FTE) who are Black and Minority Ethnic
	STU179	Number of HE students (FTE) who are Not known
	STU180	Number of HE students (FTE) who are Non-UK domicile
	STU181	Proportion of HE students (FTE) whose ethnicity is known and who are White
	STU182	Proportion of HE students (FTE) whose ethnicity is known and who are Black and Minority Ethnic
	STU184	Number of HE students (FTE) who are eligible to pay home fees
	STU185	Number of HE students (FTE) who are not eligible to pay home fees
	STU186	Number of HE students (FTE) whose eligibility to pay home fees has not been assessed
	STU187	Proportion of HE students (FTE) whose eligibility to pay home fees has been assessed and who are eligible to pay home fees
	STU189	Number of HE students (FTE) who are first year students
	STU190	Number of HE students (FTE) who are non-first year students
	STU191	Proportion of HE students (FTE) who are first year students
	STU192	Number of HE students (FTE) who are female
	STU193	Number of HE students (FTE) who are male
	STU194	Proportion of HE students (FTE) who are male
	STU195	Number of HE students (FTE) who are Postgraduate (research) students
	STU196	Number of HE students (FTE) who are Postgraduate (taught) students
	STU197	Number of HE students (FTE) who are First degree students
	STU198	Number of HE students (FTE) who are Other undergraduate students
	STU199	Proportion of HE students (FTE) who are Postgraduate (research) students
	STU200	Proportion of HE students (FTE) who are Postgraduate (taught) students
	STU201	Proportion of HE students (FTE) who are First degree students
	STU206	Number of HE students (FTE) who are "No award or financial backing" funded
	STU207	Number of HE students (FTE) who are "UK LEA mandatory/discretionary awards" funded
	STU208	Number of HE students (FTE) who are "Institutionally waived/award" funded
	STU209	Number of HE students (FTE) who are "Research councils and British Academy" funded
	STU210	Number of HE students (FTE) who are "Charities and international agencies" funded

Students	STU211	Number of HE students (FTE) who are "UK central govt/local, health, employment and agriculture authorities/bodies" funded
	STU212	Number of HE students (FTE) who are "EU sources" funded
	STU213	Number of HE students (FTE) who are "Other overseas sources" funded
	STU214	Number of HE students (FTE) who are "UK industry/commerce and students employer" funded
	STU215	Number of HE students (FTE) who are "Absent/no fees" funded
	STU216	Number of HE students (FTE) who are "Not known/Other" funded
	STU217	Proportion of HE students (FTE) whose funding method is known and who are "No award or financial backing" funded
	STU218	Proportion of HE students (FTE) whose funding method is known and who are "UK LEA mandatory/discretionary awards" funded
	STU219	Proportion of HE students (FTE) whose funding method is known and who are "Institutionally waived/award" funded
	STU220	Proportion of HE students (FTE) whose funding method is known and who are "Research councils and British Academy" funded
	STU221	Proportion of HE students (FTE) whose funding method is known and who are "Charities and international agencies" funded
	STU222	Proportion of HE students (FTE) whose funding method is known and who are "UK central govt/local, health, employment and agriculture authorities/bodies" funded
	STU223	Proportion of HE students (FTE) whose funding method is known and who are "EU sources" funded
	STU224	Proportion of HE students (FTE) whose funding method is known and who are "Other overseas sources" funded
	STU225	Proportion of HE students (FTE) whose funding method is known and who are "UK industry/commerce and students employer" funded
	STU230	Number of HE students (FTE) who are Full-time
	STU231	Number of HE students (FTE) who are Sandwich
	STU232	Number of HE students (FTE) who are Part-time
	STU233	Number of HE students (FTE) who are Writing up
	STU234	Proportion of HE students (FTE) who are not on sabbatical who are Full-time
	STU235	Proportion of HE students (FTE) who are not on sabbatical who are Sandwich
	STU236	Proportion of HE students (FTE) who are not on sabbatical who are Part-time
	STU238	Number of HE students (FTE) who are UK nationals
	STU239	Number of HE students (FTE) who are Other EU nationals
	STU240	Number of HE students (FTE) who are Non-EU nationals
	STU241	Number of HE students (FTE) whose nationality is unknown
	STU242	Proportion of HE students (FTE) whose nationality is known and who are UK nationals
	STU243	Proportion of HE students (FTE) whose nationality is known and who are Other EU nationals
	STU245	Number of HE students (FTE) who are in their foundation year
	STU246	Number of HE students (FTE) who are Year 1
	STU247	Number of HE students (FTE) who are Year 2
	STU248	Number of HE students (FTE) who are Year 3
	STU249	Number of HE students (FTE) who are Year 4
	STU250	Number of HE students (FTE) who are Year 5
	STU251	Number of HE students (FTE) who are Year 6+
	STU252	Number of HE students (FTE) who study year is unknown
	STU253	Proportion of HE students (FTE) whose year of study is known and who are in their foundation year
	STU254	Proportion of HE students (FTE) whose year of study is known and who are Year 1
	STU255	Proportion of HE students (FTE) whose year of study is known and who are Year 2
	STU256	Proportion of HE students (FTE) whose year of study is known and who are Year 3
	STU257	Proportion of HE students (FTE) whose year of study is known and who are Year 4
	STU258	Proportion of HE students (FTE) whose year of study is known and who are Year 5
Contextual Indicators	RDAP	Does the provider have Research Degree Awarding Powers
	FDAP	Does the provider have Foundation Degree Awarding Powers
	TDAP	Does the provider have Taught Degree Awarding Powers
	AOR001	Number of transnational Postgraduate research students

Aggregate Offshore Record	AOR002	Number of transnational Postgraduate taught students
	AOR003	Number of transnational First degree students
	AOR004	Number of transnational Other undergraduate students
	AOR005	Number of transnational Further education students
	AOR006	Total number of transnational students
	AOR007	Number of transnational students located within the EU
	AOR008	Number of transnational students located outside of the EU
	AOR009	Number of transnational students educated by Overseas campus of reporting HEI
	AOR010	Number of transnational students educated by Other arrangement including collaborative provision
	AOR011	Number of transnational students educated by Distance, flexible or distributed learning
	AOR012	Number of transnational students educated by Overseas partner organisation
	AOR013	Number of transnational students educated by Other arrangement
Previous Review Findings	PRV001	Outcome of previous review
	PRV003	Has ever received a negative review (1 = yes, 0 = no)
	PRV004	Outcome of previous comparable review
NSS	NSS001	Q1 - Staff are good at explaining things
	NSS002	Q2 - Staff have made the subject interesting
	NSS003	Q3 - Staff are enthusiastic about what they are teaching
	NSS004	Q4 - The course is intellectually stimulating
	NSS005	Q5 - The criteria used in marking have been clear in advance
	NSS006	Q6 - Assessment arrangements & marking have been fair
	NSS007	Q7 - Feedback on my work has been prompt
	NSS008	Q8 - I have received detailed comments on my work
	NSS009	Q9 - Feedback on my work has helped me clarify things I did not understand
	NSS010	Q10 - I have received sufficient advice & support with my studies
	NSS011	Q11 - I have been able to contact staff when I needed to
	NSS012	Q12 - Good advice was available when I needed to make study choices
	NSS013	Q13 - The timetable works efficiently as far as my activities are concerned
	NSS014	Q14 - Any changes in the course or teaching have been communicated effectively
	NSS015	Q15 - The course is well organised & is running smoothly
	NSS016	Q16 - The library resources & services are good enough for my needs
	NSS017	Q17 - I have been able to access general IT resources when I needed to
	NSS018	Q18 - I have been able to access specialised equipment, facilities or room when I needed to
	NSS019	Q19 - The course has helped me present myself with confidence
	NSS020	Q20 - My communication skills have improved
	NSS021	Q21 - As a result of the course, I feel confident in tackling unfamiliar problems
	NSS022	Q22 - Overall satisfaction
QAA Concerns	CON001	Count of QAA concerns raised, upheld or otherwise, since previous review
	CON002	Count of QAA concerns raised which do not relate to quality, standards, information or enhancement (and are therefore automatically invalid)
	CON003	Count of QAA concerns upheld since previous review which relate to academic standards
	CON004	Count of QAA concerns upheld since previous review which relate to the quality of learning opportunities
	CON005	Count of QAA concerns upheld since previous review which relate to the provision of information
	CON006	Count of QAA concerns not upheld since previous review which relate to academic standards
	CON007	Count of QAA concerns not upheld since previous review which relate to the quality of learning opportunities
	CON008	Count of QAA concerns upheld since previous review
	CON009	Count of QAA concerns not upheld since previous review which relate to the quality of learning opportunities

Table 5.10: The set of 754 metrics used in the HEI study prior to change-over-time and benchmarking calculations being added.

6. Predicting the Outcome of FEC Reviews

The purpose of this chapter is to determine which metrics, if any, could have predicted the outcome of past QAA reviews of further education colleges (FECs), and how accurately they could have done so. To do so most effectively and comprehensively two separate questions are explored:

1. Using naturally-complete metrics, could the outcome of QAA FEC reviews have been successfully predicted?
2. Using naturally-complete metrics standardised for each year, could the outcome of QAA FEC reviews have been successfully predicted?

As detailed in the earlier methods chapter, imputation would not benefit the FEC analysis and so is not considered here. The chapter begins with an overview of the HE in FE sector and its unique challenges, followed by a step-by-step description of the two analyses, which is in turn followed by a short discussion of the findings.

6.1. Introduction

Further education colleges do not have the power to award undergraduate or postgraduate degrees. Instead, they deliver higher education in one of four ways:

- Validation – a partnership arrangement where a degree-awarding body judges that a course designed and delivered by an FEC meets the requirements for a degree. Successful completion of the course results in an award from the validating body.
- Franchise arrangements – a partnership arrangement where a course designed by a degree-awarding body is delivered by the partner FEC. Successful completion of the course results in an award from the degree-awarding body.
- Higher National Certificates and Diplomas (HNCs and HNDs) – FECs teach qualifications designed and awarded by Edexcel/Pearson equivalent to the first and second year of an undergraduate degree respectively.
- Foundation Degree Awarding Powers – at the time of writing four UK colleges have been granted the right to award their own foundation degrees: degrees developed in association with employers that are equivalent to two thirds of a full honours undergraduate degree (UCAS, 2015; HEFCE, 2015b).

(McGettigan, 2013)

As the number of students at alternative providers has rocketed and the number at HEIs has continued to increase, student numbers in FECs have at best remained steady (AoC, 2013, 2014). This is in part a reflection of the localised, vocational nature of the majority of higher education provided by further education colleges (HE in FE) (Parry *et al.*, 2012). Regardless, HE in FE is provided by 254 colleges, more than the 163 traditional HEIs, and covered some 144,000 students in 2014/15 (HESA, 2015; AoC, 2014). Therefore, whilst the impact of quality in failures is small relative to HEIs, both in terms of the number of students affected and the reputation to the sector as a whole, it is still substantial in absolute terms.

The HE in FE sector shares characteristics with both the HEI and alternative provider sector. HE in FE providers are not-for-profit organisations required to report some student and finance data, although only a fraction of what is required of HEIs, to national funding councils and have, until recently, had limited incidents of 'unsatisfactory' quality assurance. As with alternative providers, the majority of HE in FE is delivered by organisations dwarfed in size by HEIs. HE in FE is unique however in that its delivery is not the main objective of the provider.

6.2. Results – Non-Standardised Metrics

Using only naturally-complete metrics, could the outcome of QAA FEC reviews have been successfully predicted?

The first FEC analysis examines which metrics, including change-over-time variants, with no missing values prior to each review could have predicted the outcome of those reviews.

6.2.1. Initial Data Exploration

To begin with a univariate analysis was run for each of the 248 metrics to determine their individual ability to predict the outcome of the reviews. By chance alone we would expect to see 248 x 0.05 = 12.4 and 248 x 0.25 = 62 significant metrics at the $p < 0.05$ and $p < 0.25$ levels respectively. This is very close to the 14 and 63 metrics found to be significant at these respective levels. The 14 metrics with a p-value < 0.05 are detailed below in Table 6.1:

Metric Code	Metric Description	P-value
FIN052_Ca1	The one-year change in net cash inflow/(outflow) from operating activities	0.004
FIN092_Ca1	The one-year change in cash generated from operations to income	0.007
FIN118_Abs	The average number of student learners per total non-teaching staff	0.012
FIN105_Abs	Dependency on higher education income	0.019
FIN100_Ca1	The one-year change in adjusted operating position (£'000)	0.025
FIN103_Cp1	The one-year percentage change in available reserves as a percentage of income	0.026
FIN103_Ca1	The one-year change in available reserves as a percentage of income	0.030
FIN056_Abs	Net cash inflow/(outflow) from returns on investments and servicing of finance	0.031
FIN107_Abs	Contribution to income from "other" income generating activities	0.033
FIN107_Ca1	The one-year change in the contribution to income from "other" income generating activities	0.034
FIN010_Ca1	The one-year change in interest and other finance costs	0.037
LEV003_Abs	Number of HE students whose level of study is 'First degree'	0.045
LEV009_Ca1	The one-year change in the number of HE students whose level of study is 'HND'	0.046
FIN083_Cp1	The one-year percentage change in total net debt	0.047

Table 6.1: All metrics from the naturally-complete FEC data set with a univariate p-value < 0.05.

The majority of significant metrics are finance related and, reassuringly, could feasibly influence the amount of resource FECs focus on quality assurance.

Examining the metric with the lowest p-value, *the one-year change in net cash inflow/(outflow) from operating activities,* it is apparent that FECs that had a reduction in net cash flow from their operating activities were subsequently more likely to be judged 'unsatisfactory' in their QAA review.

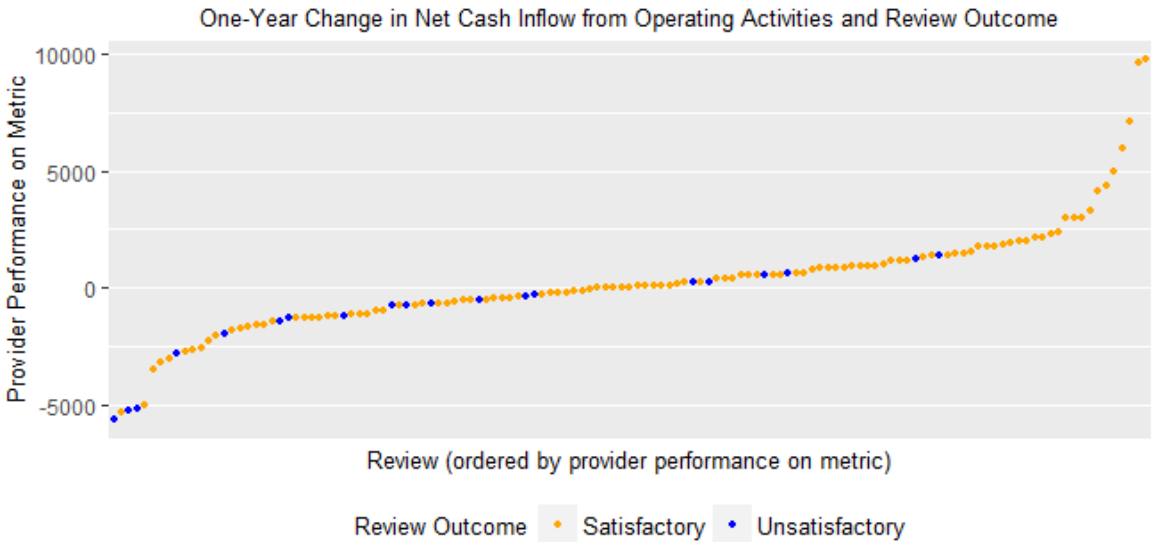

Figure 6.1: A plot of the 'one-year change in net cash inflow/(outflow) from operating activities' prior to each review and the outcome of that review.

Once again it is clear that the metric has some value. A number of 'unsatisfactory' FECs are amongst those with the greatest reduction in net cash flow from their operating activities. However, there certainly isn't a definitive relationship between a large reduction in net cash flow

from operating activities and an FEC subsequently being judged 'unsatisfactory'. There are a large number of FECs that were judged 'satisfactory' amongst the 'unsatisfactory' FECs that saw a decrease in their net cash flow from their operating activities. Moreover, there are a number of 'unsatisfactory' FECs whose performance on the metric was at or above average.

Figure 6.2 below indicates that there is a pattern for those FECs where higher education makes up a smaller proportion of their income, to be more likely to be found 'unsatisfactory'. Intuitively one can foresee that FECs where higher education is proportionally less important will be less likely, or less able, to dedicate the necessary resource to its quality assurance. Once more though the 'unsatisfactory' FECs are spread amongst a larger number of 'satisfactory' FECs. The metric does however effectively predict that those FECs for which higher education makes up a substantial proportion of their income will be 'satisfactory'.

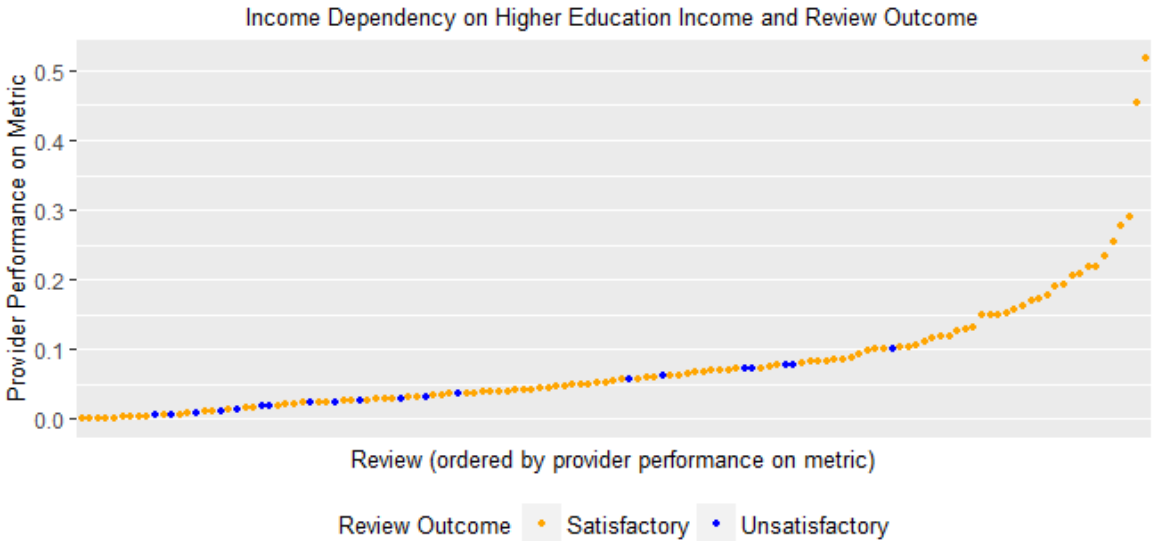

Figure 6.2: A plot of the 'income dependency on higher education income' prior to each review and the outcome of that review.

The obvious explanation for the prevalence of finance metrics exhibiting a strong relationship with the outcome of QAA FEC reviews, other than the prevalence of finance metrics in the data set, is that having 'satisfactory' quality assurance in place takes resource. If FECs are short on resource, then quality assurance activity will suffer. Another possible explanation is that the finance metrics are confounding with time. FEC budgets have come under considerable pressure in recent years (Kewin and Janowski, 2014; Wolf, 2015) which has coincided with the introduction of the more challenging *HER* methodology for FECs by the QAA (QAA, 2011a, 2014c). At the same time, as shown below in Figure 6.3, there has been a marked increase in 'unsatisfactory' judgements at FECs. The extent to which funding cuts or methodology changes are responsible for this increase in the 'unsatisfactory' performance of FECs is unclear. The impact of the sector-wide shifts over

time are minimised in the second analysis in this chapter in which metrics are standardised in-year.

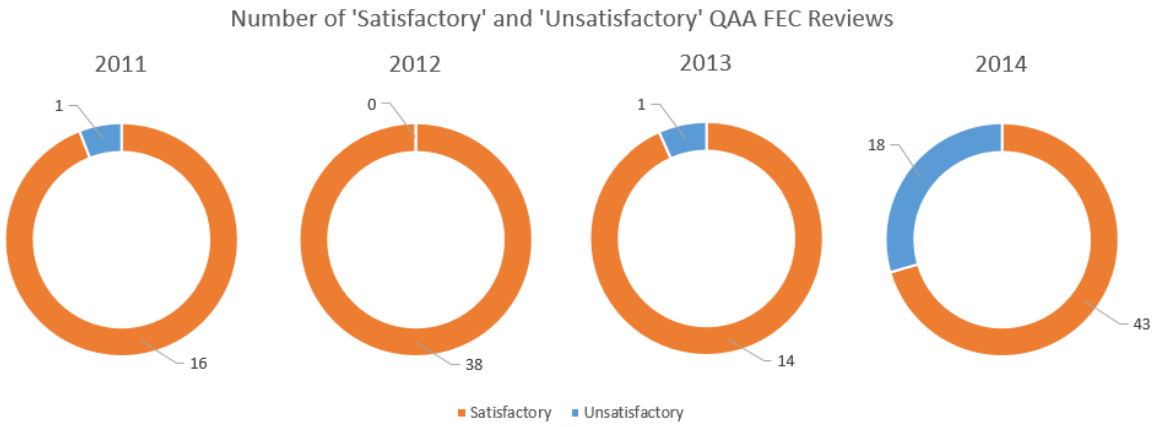

Figure 6.3: The number of 'satisfactory' and 'unsatisfactory' reviews of FECs by year.

Two metrics of interest for which the interpretation of the p-value is less straightforward due to their categorical nature are *OFS001 – Ofsted rating at the time of the QAA review* and *PRV004 - Outcome of previous comparable QAA review* shown below in Figures 6.4 and 6.5 respectively.

Ofsted reviews assess all provision, higher education or otherwise, at FECs and so, despite the small proportion of overall provision that higher education constitutes for most colleges, one might reasonably expect a strong relationship between Ofsted and QAA review outcomes. However, an FEC's Ofsted rating at the time of their review was not a significant predictor of the outcome of their QAA review. Approximately the same proportion of FECs with each Ofsted rating went on to be judged 'unsatisfactory' by the QAA, albeit a little higher for those deemed to 'require improvement' and irrelevant for those deemed 'inadequate' as low numbers – two - render such comparisons irrelevant.

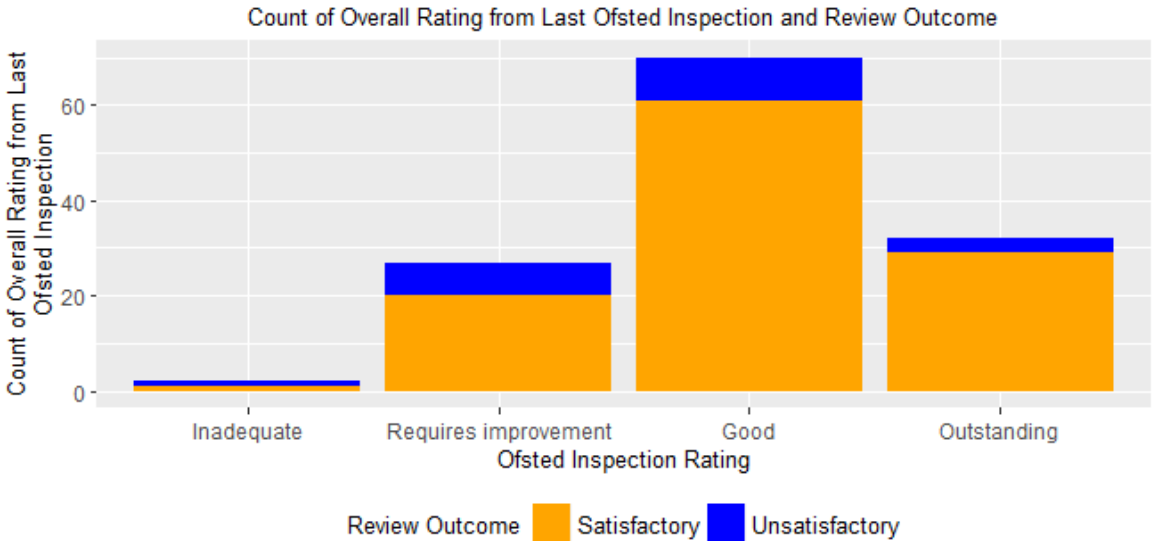

Figure 6.4: Values of *OFS001 - Ofsted rating at the time of the QAA review* and the subsequent QAA review outcome.

The 'no previous comparable review' category of the *PRV004 - Outcome of previous comparable QAA review* metric did have a significant p-value of 0.0329; however, this was because FECs that had never received a comparable QAA review were far less likely to be judged 'unsatisfactory' than FECs that had had a previous comparable review, regardless of its finding. This could simply be chance or an affectation of the cyclical nature of QAA reviews. Those who have had a previous review have simply 'had their turn' again recently facing the tougher *HER* approach and were therefore more likely to be found 'unsatisfactory'.

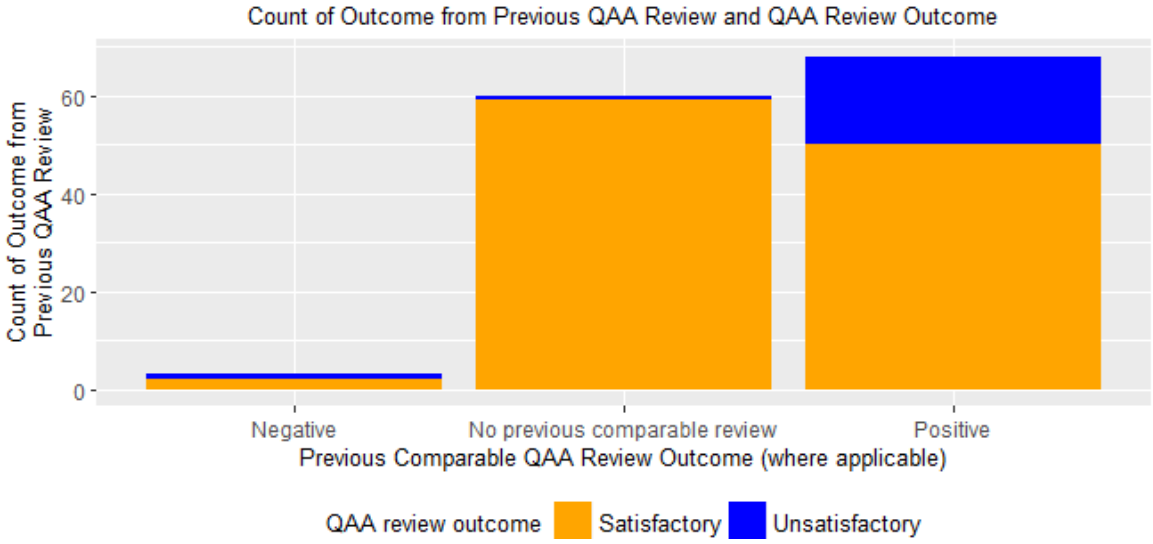

Figure 6.5: Values of *PRV004 - Outcome of previous comparable QAA review* and the subsequent QAA review outcome.

Again, the exploratory analysis of individual metrics is not reassuring. There were no more metrics with p-values of less than 0.05 or 0.25 than we would expect to see by chance alone. Moreover, these metrics exhibited a pattern familiar from the previous HEI chapter: a high proportion, but

not all, of the 'unsatisfactory' reviews were towards one end of the overall distribution, but spread amongst a large number of 'satisfactory' reviews. Those metrics one would likely have selected *a priori*, such as the outcome of Ofsted inspections or the outcome of previous QAA reviews appear to be of limited use in isolation. The next stage is to explore whether these metrics can be combined to form a useful overall model for prioritising reviews as part of a data-driven, risk-based approach to quality assurance.

6.2.2. Fitting the Model

The first stage is to develop the *elastic net* model. Running the *elastic net* procedure we obtain the diagnostic plots shown below in Figure 6.6. The λ_{min} model is again preferred to the λ_{1se} model as it is more acurate and in this case simpler too.

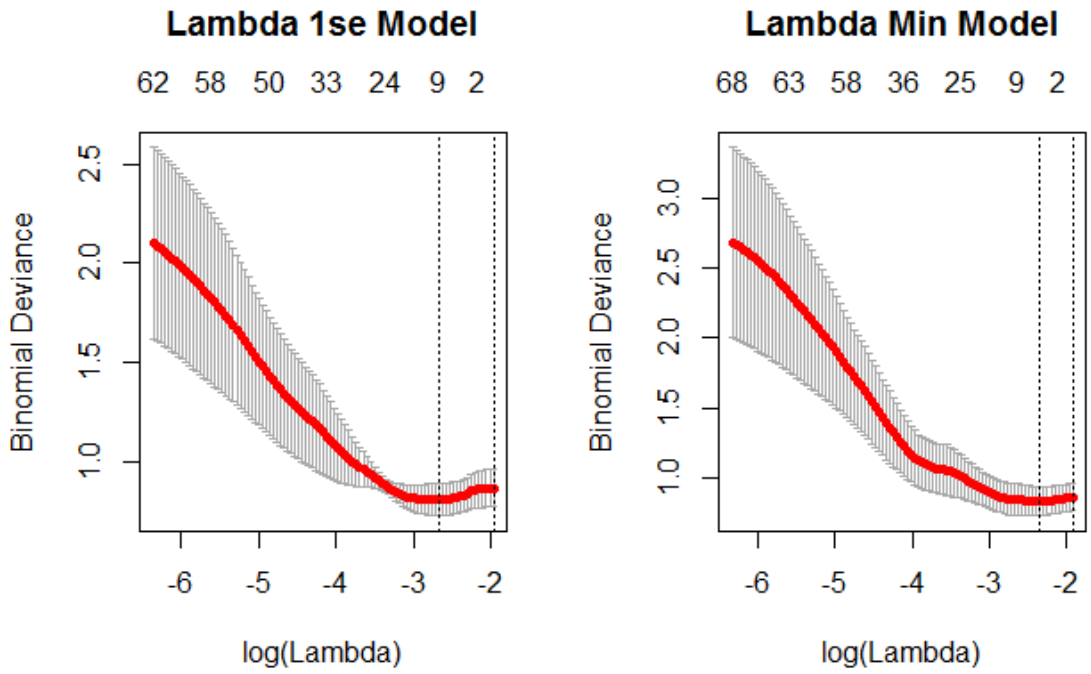

Figure 6.6: The diagnostic plots for the λ_{1se} (left) and λ_{min} (right) model for the non-standardised FEC data.

The λ_{min} model contains three metrics:

- PRV004 – the outcome of previous comparable QAA review.
- FIN107_Abs - the contribution from 'other' income generating activities. That is, non-educational income generating activities not classed as farming , catering, residences, and conferences.
- FIN103_Cp1 – the one-year percentage change in available reserves as a percentage of income.

As the *PRV004 – the outcome of previous comparable QAA review* metric has three possile values (FECs can have previously been reviewed and had a positive or negative outcome or they can have not been previously reviewed), one category – having previously received a negative review – is treated as the baseline and the model has two separate terms to account for the other eventualities. Specifically the model calculates the probability of an FEC receiving an 'unsatisfactory' review using these three metrics as follows:

$$P(Uns) = \frac{e^{(-1.67 + (-0.38 \times PRV004.NPV) + (0.16 \times PRV004.POS) + (-0.00008 \times FIN107\_Abs) + (0.14 \times FIN103\_Cp1))}}{1 + e^{(-1.67 + (-0.38 \times PRV004.NPV) + (0.16 \times PRV004.POS) + (-0.00008 \times FIN107\_Abs) + (0.14 \times FIN103\_Cp1))}}$$

Where:

PRV004.NPV = 1 if the FEC has had no previous comparable review, 0 otherwise

PRV004.POS = 1 if the FEC has had a positive previous comparable review, 0 otherwise

As the coefficient is positive for the *FIN103_Cp1 – the one-year percentage change in available reserves as a percentage of income* metric, an increase in an FEC's available reserves as a percentage of income, i.e. the amount they are relying on savings to run their day-to-day operations, will increase their predicted likelihood of being found 'unsatisfactory'. This could conceivably be a valid predictor of quality assurance review findings; those growing more reliant on reserve income will likely be looking to cut costs which could impact on their quality assurance activities. Moreover, they could feasibly already be suffering from a fall in student numbers because of their declining quality that their 'unsatisfactory' quality assurance processes have failed to halt.

As the coeffecient for the *FIN107_Abs - the contribution from 'other' income generating activities* metric is negative, the lower the contribution from 'other' income generating activities, the greater the predicted likelihood of being found 'unsatisfactory'. As FIN017_Abs is in absolute metric, it will make no difference how much this value has changed since the previous year or what percentage of of income generating activities are categorised as 'other' – only the absolute size of the income generated which has been categorised as 'other'. It is hard to see any obvious meaningful connection between 'other' income generating activities and the outcome of quality assurance reviews.

An FEC that has a positive previous comparable QAA review will, somewhat counterintuitively, see its predicted likelihood of being found 'unsatisfactory' increase whilst the oppostive is true for an FEC with no previous comparable review. As discussed earlier, this is likely to be a proxy for the timing of the review and thus whether or not the organisation has undergone a tougher *HER* review. There is however also the possibility that FECs previously judged 'satisfactory' may be

more likely to become complacent with regards their quality assurance processes in contrast to those FECs that have previously been found 'unsatisfactory' and therefore are focusing on improvement.

6.2.2.1. A Worked Exploration of the Model

To better understand the effect of changes in performance for each metric contained in the model we can consider three hyopthetical FECs whose performance, and the resulting predicted probability of being judged 'unsatisfactory', are detailed below.

First, if we consider three hypothetical FECs all of which had neither generated nor lost 'other' income in the most recent financial year and had no change in the percentage of their income that derived from available reserves such that FIN107_Abs = 0 and FIN103_Cp1 = 0 for all three. Furthermore, FEC A has had a previous review which resulted in an 'unsatisfactory' judgement, FEC B has had no previous review, and FEC C has had a previous review which resulted in a 'satisfactory' judgement. Their performance and the resulting predicted likelihood of being judged 'unsatisfactory' can be seen below in Table 6.2:

FEC	PRV004.NPV	PRV004.POS	Predicted Probability of 'Unsatisfactory' Judgement
A	0	0	$\frac{e^{(-1.67 + (-0.38\times0)+(0.16\times0)+(-0.00008\times35)+(-0.14\times0))}}{1 + e^{(-1.67 + (-0.38\times0)+(0.16\times0)+(-0.00008\times35)+(-0.14\times0))}} = 15.80\%$
B	1	0	$\frac{e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times35)+(-0.14\times0))}}{1 + e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times35)+(-0.14\times0))}} = 11.38\%$
C	0	1	$\frac{e^{(-1.67 + (-0.38\times0)+(0.16\times1)+(-0.00008\times35)+(-0.14\times0))}}{1 + e^{(-1.67 + (-0.38\times0)+(0.16\times1)+(-0.00008\times35)+(-0.14\times0))}} = 18.05\%$

Table 6.2: The predicted likelihoods of hypothetical FECs being judged 'unsatisfactory'.

Having previously had a postive QAA review gives FEC C the highest predicted likelihood of being unsatisfactory (18.05%), FEC A has the second highest predicted probability of being judged 'unsatisfactory' (15.80%) having previously had a negative QAA review, and finally FEC B which has never had a QAA review has the lowest predicted probability of being judged 'unsatisfactory' (11.38%).

Second, we consider three FECs each of which has had no previous comparable QAA review and had no change in the percentage of their income that derived from available reserves such that PRV004.NPV = 1, PRV004.POS = 0, and FIN103_Cp1 = 0 for all three. Furthermore, FEC A has an average contribution from 'other' income generating activities of £35,000, FEC B has income from 'other' contributions of £5,000,000 (towards the maximum in the data set), and FEC C has a loss of £5,000,000 from 'other' activities (towards the maximum in the data set). Their performance

and the resulting predicted likelihood of being judged 'unsatisfactory' can be seen below in Table 6.3:

FEC	FIN107_Abs	Predicted Probability of 'Unsatisfactory' Judgement
A	35	$\frac{e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times35)+(-0.14\times0))}}{1 + e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times35)+(-0.14\times0))}} = 11.38\%$
B	5,000	$\frac{e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times5000)+(-0.14\times0))}}{1 + e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times5000)+(-0.14\times0))}} = 7.94\%$
C	-5,000	$\frac{e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times-5000)+(-0.14\times0))}}{1 + e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times-5000)+(-0.14\times0))}} = 16.11\%$

Table 6.3: The predicted likelihoods of hypothetical FECs being judged 'unsatisfactory'.

Using some of the more extreme entries in the data set the FIN017_Abs metric has a slightly greater impact than an FECs previous QAA review activity.

Third, we consider three FECs each of which has had no previous comparable QAA review and neither lost nor generated 'other' income such that PRV004.NPV = 1, PRV004.POS = 0, and FIN107_Abs = 0 for all three. Furthermore, FEC A has an average one-year change in available reserves as a percentage of income of 13.26% while FEC B and FEC C have changes of -100% and 100%, towards the extremes of entries in the data set, respectively. Their performance and the resulting predicted likelihood of being judged 'unsatisfactory' can be seen below in Table 6.4:

FEC	FIN103_Cp1	Predicted Probability of 'Unsatisfactory' Judgement
A	13.26%	$\frac{e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times0)+(0.14\times0.1326))}}{1 + e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times0)+(0.14\times0.1326))}} = 11.22\%$
B	-100.00%	$\frac{e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times0)+(0.14\times-1))}}{1 + e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times0)+(0.14\times-1))}} = 10.07\%$
C	100.00%	$\frac{e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times0)+(0.14\times1))}}{1+ e^{(-1.67 + (-0.38\times1)+(0.16\times0)+(-0.00008\times0)+(0.14\times1))}} = 12.90\%$

Table 6.4: The predicted likelihoods of hypothetical FECs being judged 'unsatisfactory'.

The impact of the metric when considered in isolation and with extreme values is less than for the other two metrics contained in the model.

By combining the three hypothetical FECs performance when the metrics were allowed to vary we can see the overall output from the model in Table 6.5 below.

FEC	PRV004.NPV	PRV004.POS	FIN107_Abs	FIN103_cp1	Predicted Probability of 'Unsatisfactory' Judgement
A	0	0	35	13.26%	15.56%
B	1	0	5,000	-100.00%	6.98%
C	0	1	-5,000	100.00%	27.49%

Table 6.5: The predicted likelihoods of hypothetical FECs being judged 'unsatisfactory'.

FEC B, performing in a manner that leads to each metric in the model lowering its predicted probability of being judged 'unsatisfactory' has a predicted probability of being judged 'unsatisfactory' of 6.98%. FEC C, performing in a manner that leads to each metric raising its predicted probability of being judged 'unsatisfactory' has a predicted probability of being judged 'unsatisfactory' of 27.49% - 3.94 times greater than FEC B.

As with the metrics considered in isolation, it is clear that the effects of changes in performance on the predicted likelihood of being judged 'unsatisfactory' can be substantial in relative terms, but is still fairly limited in absolute terms. The narrow range of predicted probabilities once more suggests that the model will have a high error rate and, as with the HEI model, is indicative of the lack of a strong relationship between the metrics and QAA review outcomes. The fact that the model cannot identify which FECs will or will not be found 'unsatisfactory' with any degree of certainty has again resulted in the predicted likelihoods of the event occuring clustering close to the probability of being judged 'unsatisfactory' regardless of the data, in this instance 20 ÷ 131 = 15.27%. That is, all FECs will be predicted to be more or less equally likely to be 'unsatisfactory' regardless of thei performance.

6.2.3. Evaluating the Model

6.2.3.1. Testing the Fit of the Model

Figure 6.7 below shows the ROC curve for this model when applied to the data used to develop it. The fairly impressive 'area under the curve' value of 0.854 suggests a reasonable rate of 'unsatisfactory' FECs being succesfully prioritised as the threshold criteria for triggering a review is lowered:

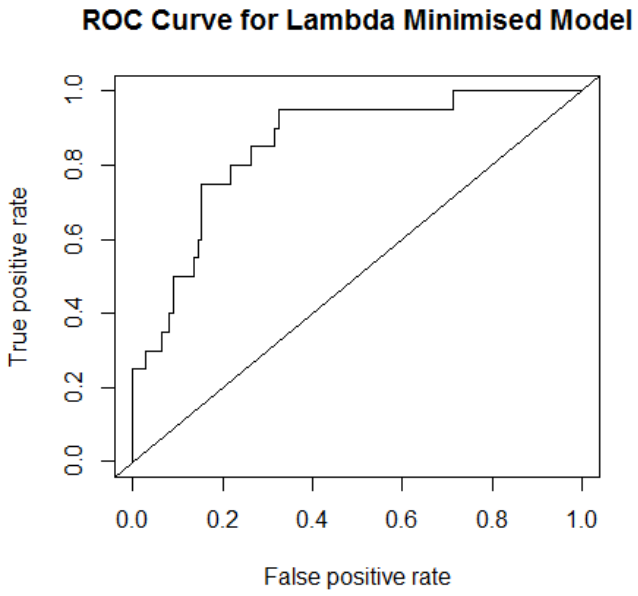

Figure 6.7: The ROC curve for the FEC model featuring the non-standardised metrics PRV004, FIN107_Abs and FIN103_Cp1.

Table 6.6 and Figure 6.8 below show in greater detail the effect of lowering the threshold required to trigger a review. The model's performance appears an improvement on what was achieved for HEIs. The first five 'unsatisfactory' reviews could have been prioritised without any 'satisfactory' reviews being incorrectly prompted and, at the optimal cuf-off point (statistically speaking), 17 'unsatisfactory' reviews (three short of the total of 20) could have been prioritised with only 27 'satisfactory' reviews being incorrectly prioritised. To have prioritised all 'unsatisfactory' reviews 102 reviews would have been required – an error rate of 82 / (20 + 82) = 80.39% amongst those FECs prioritised for review. This is an improvement on the error rate of 88.79% obtained for HEIs when the one hard-to-predict 'unsatisfactory' review was discounted; however, it is still means more than four in five FECs prioritised for review by the model would have been judged 'satisfactory'. This high error rate is unlikely to sit well with FECs prioritised for review by the model.

← Decreasing probability of a review being 'unsatisfactory' required to trigger inspection	Predicted Probability of an 'unsatisfactory' outcome required to trigger a review	Number of 'unsatisfactory' reviews (true positives)	Number of 'satisfactory' reviews (false positives)	Accuracy rate
	45.05%	1	0	0.85
	28.81%	2	0	0.86
	25.10%	3	0	0.87
	23.09%	4	0	0.88
	20.49%	5	0	0.89
	18.70%	6	3	0.87
	18.48%	7	5	0.86
	18.45%	8	8	0.85
	18.39%	9	9	0.85
	18.36%	10	9	0.85
	18.35%	11	9	0.86
	18.24%	12	15	0.82
	18.22%	13	16	0.82
	18.18%	14	16	0.83
	18.16%	15	16	0.84
	18.02%	16	27	0.76
	18.01%	17	27	0.77
	17.85%	18	34	0.73
	15.77%	19	49	0.62
	11.39%	20	82	0.37

Table 6.6: The number of 'satisfactory' and 'unsatisfactory' FEC reviews that would have resulted from decreasing the threshold required to prompt a review using the λ_{min} model.

It is clear from Figure 6.8 there is an initial cluster of succesful predictions followed by an area of reasonable success before the number of incorrectly prioritsed 'satisfactory' reviews begins to increase. It appears there is a set of characteristics shared by the majority – although certainly not all – 'unsatisfactory' FECs, along with a greater number of 'satisfactory' FECs. As noted in the earlier worked example of the model, there is a narrow range of predicted probabilities which contributes to a high error rate. No FEC is predicted to be 'unsatisfactory' with a probability greater than 45% and the majority of FECs have very similar predicted probabilities of being 'unsatisfactory'.

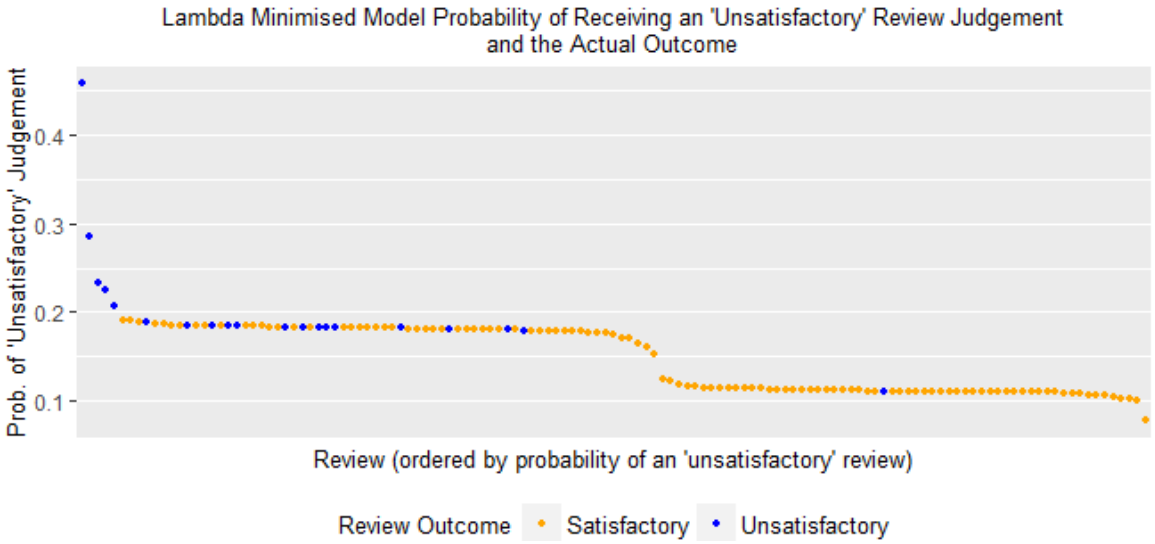

Figure 6.8: Predicted probabilities for each of the 131 complete, comparable reviews and their actual outcome.

Setting aside the narrow range of predicted probabilities and focusing on the ordering of reviews, the model, at first, does a reasonable job of prioritising 'unsatisfactory' providers. The final three 'unsatisfactory' FECs however mean that, if no 'unsatisfactory' provision is deemed acceptable, nearly all FECs would need to be reviewed. However, the low number of metrics in the model suggests that it is not overfitting the data; the early success and low number of metrics in the model indicates that it is picking-up on a genuine signal in the data.

There are two further tests we can perform at this point to assess the accuracy and usefulness of the model. *First,* we can look at the performance of the model's predictions at a specific point in the past to assess its real world use. *Second*, we can look to see how the model performed at predicting the outcome of the reviews that have taken place since the original data set for this study was collected. This will show how the model would have performed had it been used by the QAA after September 2014.

6.2.3.2. Assessing the Model's Predictions

Figure 6.9 below shows the distribution of the predicted probabilities of each FEC being 'unsatisfactory' on 5th November 2013. Those points coloured white represent FECs not reviewed within a year of 5th November 2013, those coloured orange represent those FECs found 'satisfactory' within one year, and those points coloured blue represent those found 'unsatisfactory' within one year.

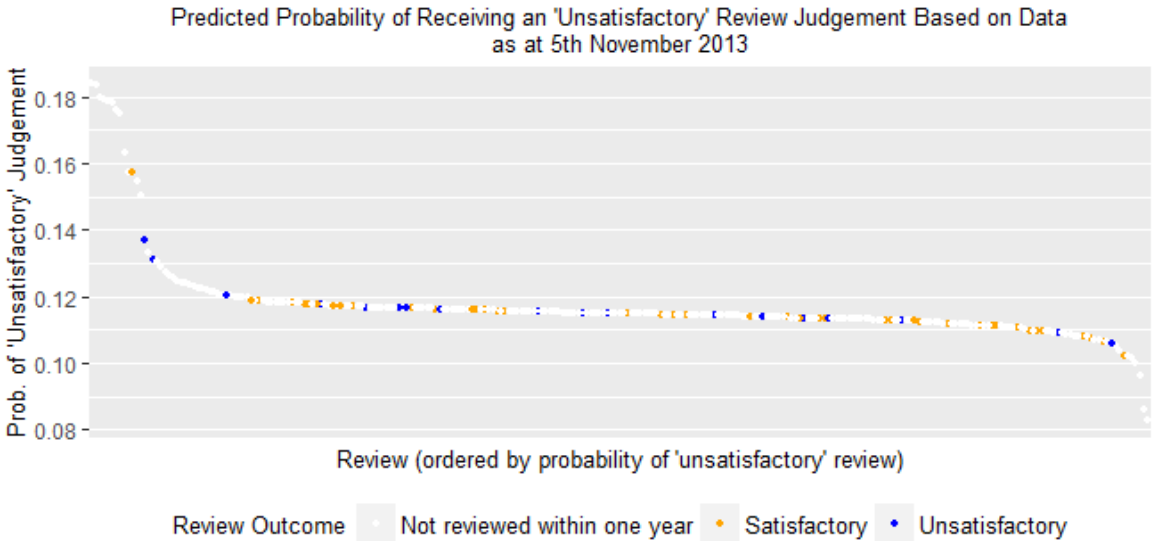

Figure 6.9: Predicted probabilities of each FEC being 'unsatisfactory' on 5th November 2013.

One can see that there were a sizeable number of FECs deemed at greater risk of being 'unsatisfactory' than those that were reviewed as part of the QAA's cyclical approach. Figure 6.10 below shows the same data with the FECs not reviewed within one year removed for a clearer picture:

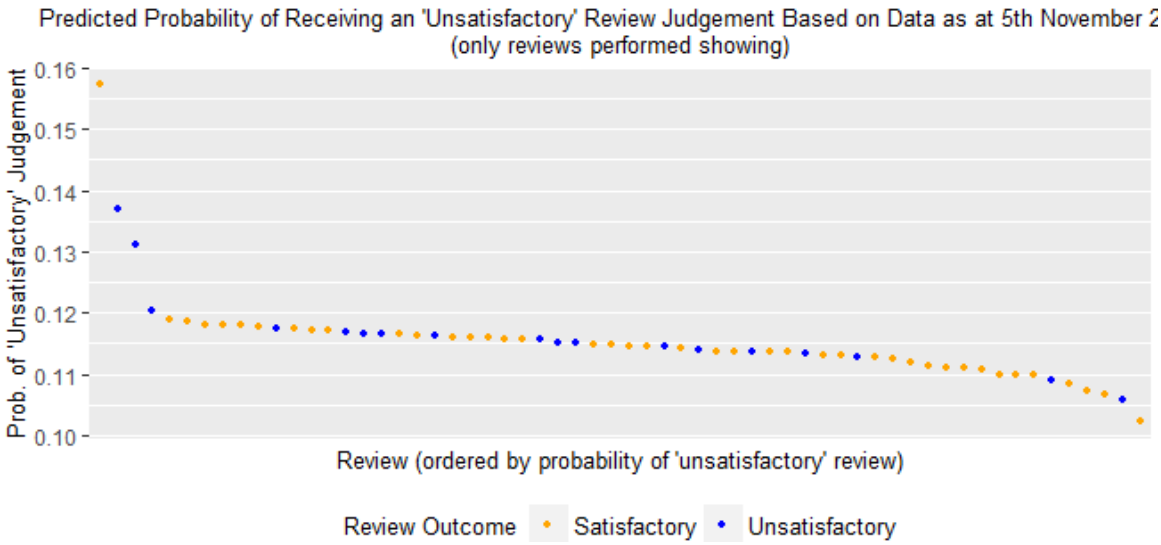

Figure 6.10: Predicted probabilities of each FEC being 'unsatisfactory' on 5th November 2013 with those FECs not reviewed within one year removed.

The very narrow range of predicted probabilities, as indicated by the shallow gradient of most of the curve, and, of greater concern, the significant number of 'unsatisfactory' FECs towards the 'lower risk' end of the distribution suggest the model would not have been succesful.

Figure 6.11 below show the model's application to all 50 of the QAA FEC reviews which took place from September 2014 to July 2015 after the initial data set for this study was obtained.

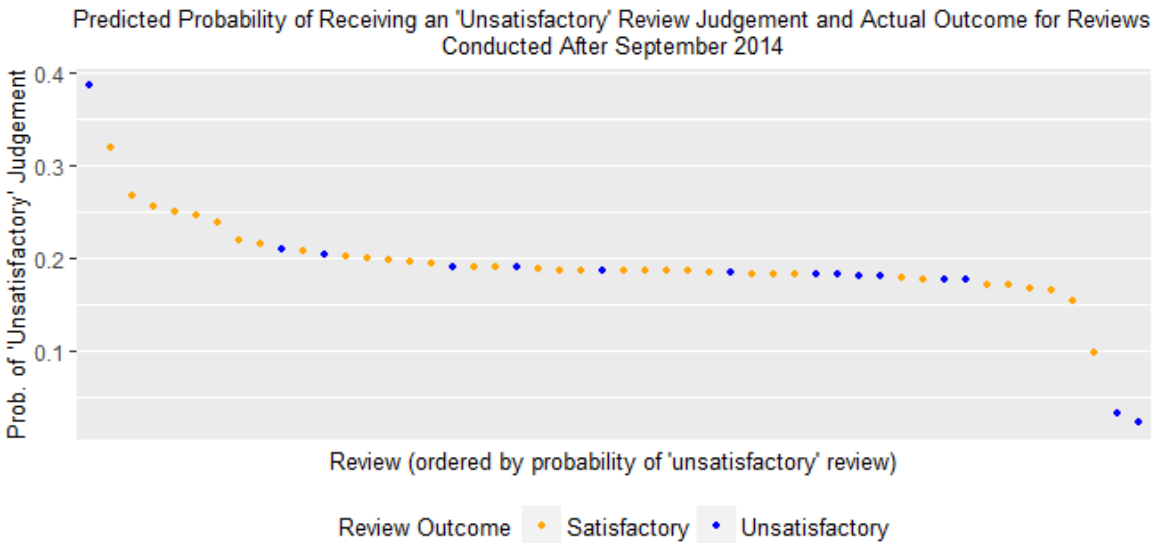

Figure 6.11: The predicted probabilities and actual outcomes of FEC reviews conducted since the original data set was obtained and therefore not used in the development of the model.

The model performs poorly: nine of the 14 'unsatisfactory' FECs were amongst the bottom half of FECs prioritised.

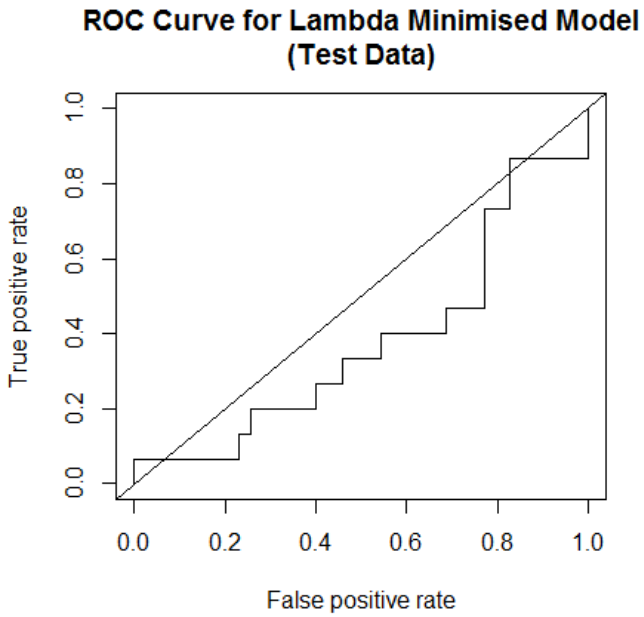

Figure 6.12: The ROC curve for the model applied to FEC reviews conducted since the original data set was obtained and therefore not used in the development of the model.

The area under the ROC curve is 0.379 which, as it is less than 0.5, confirms that the QAA would be better off doing the exact opposite of what the model suggests and prioritising for review those predicted as being *least* likely to be 'unsatisfactory' and working their way through to those predicted *most* likely. No one realistically expects a data-driven, risk-based system to perfectly predict those providers that will be 'unsatisfactory', but at an absolute minimum a greater

proportion of those providers deemed high risk should be found to be 'unsatisfactory' compared to those deemed low risk. This model cannot do that.

6.2.4. Summary

For this question, we have considered all metrics with a feasible link to quality assurance that could form part of a cost-effective, data-driven, risk-based approach, not just in their absolute state but also modified to account for changes over time, both in percentage and absolute terms. The best predictive model was determined to be:

$$P(Uns) = \frac{e^{(-1.67 + (-0.38 \times PRV004.NPV) + (0.16 \times PRV004.POS) + (-0.00008 \times FIN107\_Abs) + (0.14 \times FIN103\_Cp1))}}{1 + e^{(-1.67 + (-0.38 \times PRV004.NPV) + (0.16 \times PRV004.POS) + (-0.00008 \times FIN107\_Abs) + (0.14 \times FIN103\_Cp1))}}$$

Where:

- FIN107_Abs is the contribution from 'other' income generating activities
- FIN103_Cp1 is the one-year percentage change in available reserves as a percentage of income
- PRV004.NPV = 1 if the FEC has had no previous comparable review, 0 otherwise
- PRV004.POS = 1 if the FEC has had a positive previous comparable review, 0 otherwise

However, the answer to the question:

Using naturally-complete metrics, could the outcome of QAA FEC reviews have been successfully predicted?

is no. The model initially appeared an improvement on the HEI model with just 16 'satisfactory' FECs being unnecessarily prioritised for review by the point 15 'unsatisfactory' FECs, three-quarters of the total, had been correctly prioritised. However, to have prioritised all 'unsatisfactory' FECs for review would have required 82 'satisfactory' FECs to have been unnecessarily reviewed: an error rate of 80.4% amongst those FECs prioritised for review, and sparing just 29 of the 111 'satisfactory' FECs in the data set from review.

Some may argue that by allowing five 'unsatisfactory' FECs to go unreviewed all but 16 of the 'satisfactory' FECs could be spared a review and the model is therefore a success. Such a result, however, relies on knowing exactly when to stop conducting reviews, a difficult decision without the benefit of perfect hindsight, and ignores the more realistic application of the model to a specific point in time. The 'unsatisfactory' FECs are distributed evenly throughout and had the QAA conducted their reviews in order of the predicted probability that they would result in an 'unsatisfactory' judgement, a large number of 'satisfactory' FECs would have been reviewed and a large number of 'unsatisfactory' FECs would not have been reviewed.

Of greater concern are the narrow range of predicted probabilities and the application of the model to new data. In this case the model is actively misleading suggesting that either the model was based on chance relations despite the efforts taken to guard against this, or that the weak underlying relationship between metrics and review outcomes which did exist no longer does. The next step is to see if the accuracy and continued use of the model can be improved by taking into account changes in the sector over time with the use of in-year standardised metrics.

6.3. Results – Standardised Metrics

Using naturally-complete, in-year standardised metrics, could the outcome of QAA FEC reviews have been successfully predicted?

6.3.1. Initial Data Exploration

From the initial univariate analyses there were a greater number of standardised metrics with a p-value less than 0.05 than was the case for the unstandardised metrics. These metrics are similar in nature to the significant non-standardised metrics focusing mainly on finance.

Metric Code	Metric Description	P-value
FIN057_Abs	Cash Flow Statement - Taxation	0.000007
CON002	Count of QAA concerns raised which do not relate to quality, standards, information or enhancement (and are therefore automatically invalid)	0.00011
PRV013	Worst judgement concerning enhancement in previous comparable review	0.000138
FIN010_Ca1	Income and Expenditure - Expenditure - Interest and other finance costs	0.009946
PRV011	Worst judgement concerning information in previous comparable review	0.014564
FIN107_Ca1	Ratio analysis - Income generating activities - Contribution from other income generating activities	0.014895
FIN103_Ca1	Ratio analysis - Margin - Available reserves as a percentage of income	0.018166
FIN105_Abs	Ratio analysis - Income - Dependency on Higher Education income	0.025443
FIN092_Ca1	Ratio analysis - Cash generated from operations to income	0.029809
FIN118_Abs	Supplementary Benchmarking Information - Expenditure -Staff number and cost details and staff numbers/costs per SLN/FTE - Ave SLN learners per total non-teaching staff	0.034173
LEV008_Abs	Proportion of HE students whose level of study is 'HNC'	0.034651
FIN107_Abs	Ratio analysis - Income generating activities - Contribution from other income generating activities	0.036317
FIN056_Abs	Cash Flow Statement - Returns on investments and servicing of finance - Net cash inflow/(outflow) from returns on investments and servicing of finance	0.040083
FIN067_Ca1	Cash Flow Statement - Financing - Repayment of amounts borrowed - secured and unsecured loans	0.040694
FIN056_Ca1	Cash Flow Statement - Returns on investments and servicing of finance - Net cash inflow/(outflow) from returns on investments and servicing of finance	0.041048
AGE001_Abs	Number of HE students aged under 21	0.042613
LEV003_Abs	Number of HE students whose level of study is 'First degree'	0.042787
FIN093_Ca1	Ratio analysis - Gearing - Debt charges as a percentage of income	0.045885
FIN049_Ca1	Balance Sheet - I&E account including pension reserve	0.045921
LEV008_Ca1	Proportion of HE students whose level of study is 'HNC'	0.046601

Table 6.7: All metrics from the standardised FEC data set with a univariate p-value < 0.05.

Looking at the metric with the lowest p-value *FIN057_Abs - Cash Flow Statement – Taxation* we can see a quirk of the standardisation process. One extreme outlier can have the effect of, for all intents and purposes, reducing the genuine, lesser variation between other providers down to zero.

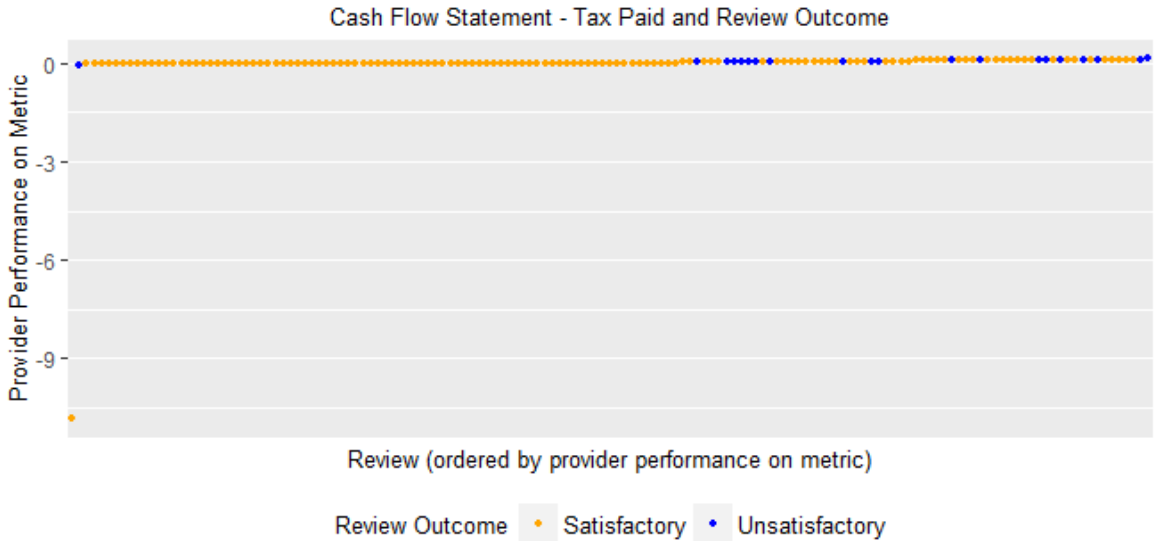

Figure 6.13: A plot of the standardised 'Cash Flow Statement – Taxation' metric prior to each review and the outcome of that review.

This problem could in theory be rectified by using Winsorisation where, in its simplest form, cases are ranked according to their naive z-scores and the top and bottom 10% for example are all assigned the z-score of the FEC at the 10th and 90th percentile respectively (Spiegelhalter, 2005). Whilst this reduces the impact of outliers it does so by altering some, in this example 20%, of providers' performance data masking their poor performance and with it making interpretation of the metrics more challenging. It is therefore not used here. In the case of the *FIN057_Abs - Cash Flow Statement – Taxation* metric, the p-value is significant because the single extreme value relates to a 'satisfactory' review and hence extreme negative values are perfectly associated with being subsequently judged 'satisfactory'. Fortunately, the majority of metrics are not affected by extreme outliers.

The *LEV003_Abs – Number of HE students whose level of study is 'first degree'* (i.e. an undergraduate degree) metric shown below in Figure 6.14 is more promising. Whilst it is again the case that the 'unsatisfactory' FECs are distributed amongst a number of 'satisfactory' FECs, those FECs with an above average number of students studying for a first degree, i.e. with a z-score greater than zero, are all 'satisfactory'. Whilst metrics such as this cannot pinpoint 'unsatisfactory' provision, they can possibly help avoid prioritising 'satisfactory' FECs for review.

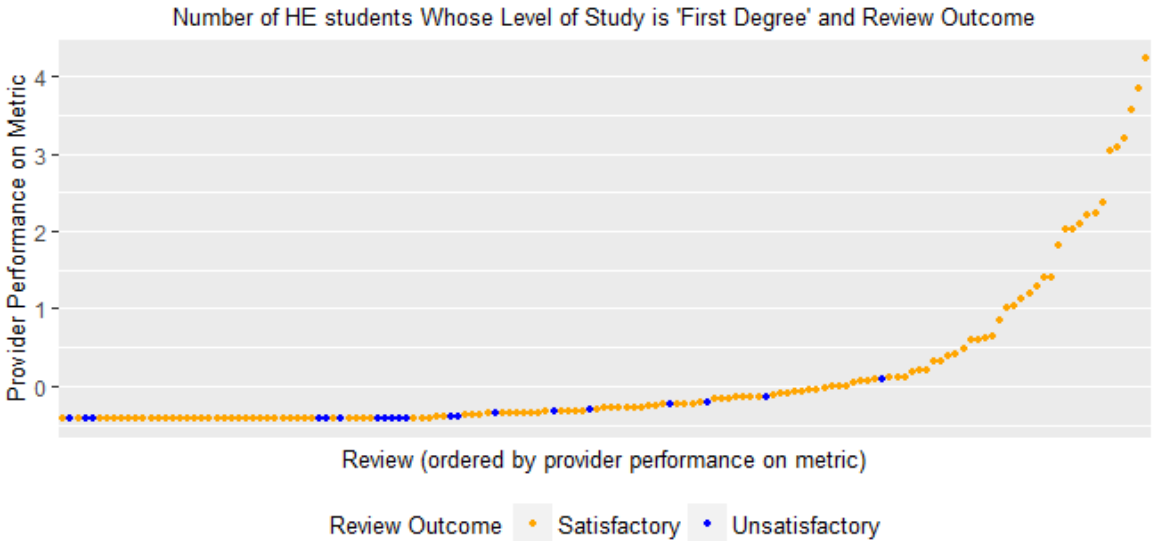

Figure 6.14: A plot of the standardised 'Proportion of HE students whose level of study is HNC' metric prior to each review and the outcome of that review.

6.3.2. Fitting the Model

When attempting to combine these in-year standardised metrics into a predictive model the result is that, as with HEIs, no cross-validated model performs significantly better than simply treating all FECs as equally likely to be judged 'unsatisfactory' regardless of the available data.

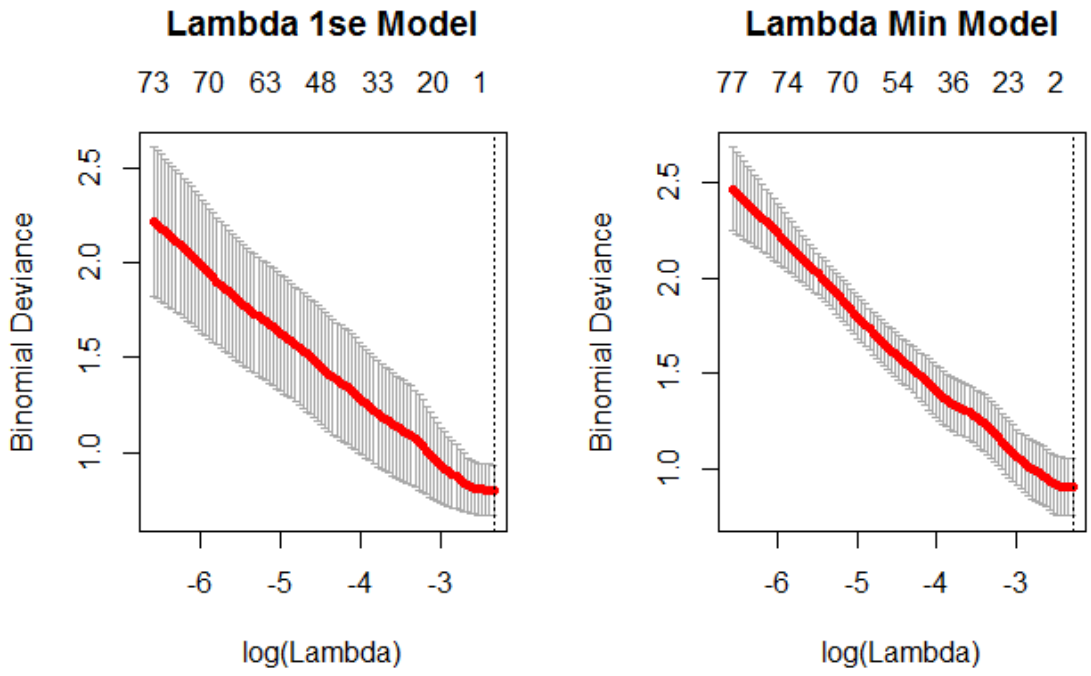

Figure 6.15: The diagnostic plots for the λ_{1se} (left) and λ_{min} (right) model for standardised FEC data.

Therefore, using naturally-complete, in-year standardised metrics, the outcome of QAA FEC could not have been successfully predicted. It appears that standardising the data within year has in fact hindered, rather than helped, with the prediction of 'unsatisfactory' FECs. The most likely

suggestion as for why this might be is the same as for HEIs: that review outcomes are absolute not relative. If, for example, an FEC is in dire financial straits, it's quality assurance processes will not be helped by other FECs being in an even worse financial position.

6.4. Summary

The best model calculated the predicted probability of an FEC being 'unsatisfactory' using non-standardised versions of the outcome of the previous comparable QAA review, the contribution from 'other' income generating activities, and the one-year percentage change in available reserves as a percentage of income. Specifically the model was defined as:

$$P(Uns) = \frac{e^{(-1.67 + (-0.38 \times PRV004.NPV) + (0.16 \times PRV004.POS) + (-0.00008 \times FIN107\_Abs) + (0.14 \times FIN103\_Cp1))}}{1 + e^{(-1.67 + (-0.38 \times PRV004.NPV) + (0.16 \times PRV004.POS) + (-0.00008 \times FIN107\_Abs) + (0.14 \times FIN103\_Cp1))}}$$

Where:

PRV004.NPV = 1 if the FEC has had no previous comparable review, 0 otherwise

PRV004.POS = 1 if the FEC has had a positive previous comparable review, 0 otherwise

In seeking to determine which metrics, if any, would have successfully predicted the outcome of QAA FEC reviews we have considered all metrics with a feasible link to quality assurance that could form part of a cost-effective, data-driven, risk-based approach, not just in their absolute state but also modified to account for changes over time, both in percentage and absolute terms. The best predictive model performed well to begin with but had a high error rate and its predictions were very poor when applied retrospectively to a specific point in time and to new reviews. Indeed, the QAA would have successfully prioritised more 'unsatisfactory' new reviews by focusing on those predicted as being *least* likely to be 'satisfactory'. Standardising the metrics to measure deviations from the norm for the year in which the data was recorded resulted in no model being a significant improvement on assuming an equal likelihood for all FECs being judged 'unsatisfactory' and ignoring the data entirely. Therefore, neither standardised nor non-standardised metrics could have successfully predicted the outcome of QAA FEC reviews.

6.5. Discussion

As with HEIs, no effective model would have allowed the QAA to successfully prioritise 'unsatisfactory' FEC reviews. Even if one loosely defines success as identifying a 'high risk' group of FECs containing a greater proportion of 'unsatisfactory' FECs than a 'low risk' group – a

definition the QAA and students would be unlikely to accept –the model could not have achieved this at a specific point in time.

The FEC model was the only one that could be tested to see how it performed with new data and new reviews following the development of the model. That performance was bad. The inability of the model to effectively predict the the outcome of a new set of QAA FEC reviews suggests that the weak links between the metrics and review outcomes no longer holds (if indeed they were ever more than coincidental). This highlights the challenges of using any predictive measures in a constantly changing environment. What is a succesful predictor one day may no longer be the next. The severity of this problem will be environment specific: there have not been regular changes requiring Amazon to make substantive changes to its model for predicting what customers might like based on past purchases, nor have there been substantive changes requiring the Met Office to change the factors considered when forecasting weather. The higher education sector is different however: devolution, funding cuts to FECs, the marketisation of the sector, the tripling of tuition fees, the removal of the student numbers cap, student visa changes and Brexit could all feasibly change the sector such that existing relationships between metrics and the outcome of QAA reviews breakdown and new ones are formed (or not). This topic will be discussed in depth in chapter eight which explores the reasons behind the quantitative finding.

It was suggested in the discussion of the previous HEI chapter that, arguably, despite not being possible for HEIs, operating a purely data-driven, risk-based approach to prioritising QAA reviews could still prove useful if it could succesfully prioritise FEC and alternative provider reviews. This chapter has demonstrated that such a risk-based approach is also not viable for FECs. The one hope left for such an approach is for alternative providers which, although they only make up a fraction of the sector by student numbers, are the most numerous and subject to the greatest number of reviews.

Appendix F – FEC Metrics

The set of 181 metrics used in this study prior to change-over-time calculations being added.

Area	Metric Code	Metric Description
Student Characteristics	AGE001	Number of HE students aged under 21
	AGE002	Proportion of HE students aged under 21
	AGE003	Number of HE students aged 21-24
	AGE004	Proportion of HE students aged 21-24
	AGE005	Number of HE students aged 25 and over
	AGE006	Proportion of HE students aged 25 and over
	DOM001	Number of HE students whose domicile is 'Other EU'
	DOM002	Proportion of HE students whose domicile is 'Other EU'
	DOM003	Number of HE students whose domicile is 'non-EU'
	DOM004	Proportion of HE students whose domicile is 'non-EU'
	DOM005	Number of HE students whose domicile is 'UK'
	DOM006	Proportion of HE students whose domicile is 'UK'
	ENT001	Number of HE students who are entrants
	ENT002	Proportion of HE students who are entrants
	ENT003	Number of HE students who are non-entrants
	ENT004	Proportion of HE students who are non-entrants
	ETH001	Number of HE students who are Asian or Asian British
	ETH002	Proportion of HE students who are Asian or Asian British
	ETH003	Number of HE students who are Black or Black British
	ETH004	Proportion of HE students who are Black or Black British
	ETH005	Number of HE students whose ethnicity is 'Other (including mixed)'
	ETH006	Proportion of HE students whose ethnicity is 'Other (including mixed)'
	ETH007	Number of HE students whose ethnicity is unknown
	ETH008	Proportion of HE students whose ethnicity is unknown
	ETH009	Number of HE students who are White
	ETH010	Proportion of HE students who are White
	GEN001	Number of HE students who are female
	GEN002	Proportion of HE students who are female
	GEN003	Number of HE students who are male
	GEN004	Proportion of HE students who are male
	LEV001	Number of HE students whose level of study is 'Diploma'
	LEV002	Proportion of HE students whose level of study is 'Diploma'
	LEV003	Number of HE students whose level of study is 'First degree'
	LEV004	Proportion of HE students whose level of study is 'First degree'
	LEV005	Number of HE students whose level of study is 'Foundation degree'
	LEV006	Proportion of HE students whose level of study is 'Foundation degree'
	LEV007	Number of HE students whose level of study is 'HNC'
	LEV008	Proportion of HE students whose level of study is 'HNC'
	LEV009	Number of HE students whose level of study is 'HND'
	LEV010	Proportion of HE students whose level of study is 'HND'

Student Characteristics	LEV011	Number of HE students whose level of study is 'Postgraduate research'
	LEV012	Proportion of HE students whose level of study is 'Postgraduate research'
	LEV013	Number of HE students whose level of study is 'Postgraduate taught'
	LEV014	Proportion of HE students whose level of study is 'Postgraduate taught'
	LEV015	Number of HE students whose level of study is 'Undergraduate other'
	LEV016	Proportion of HE students whose level of study is 'Undergraduate other'
	MOD001	Number of HE students who are full-time
	MOD002	Proportion of HE students who are full-time
	MOD003	Number of HE students who are part-time
	MOD004	Proportion of HE students who are part-time
Finance	FIN001	Income and Expenditure - Income - Funding body grants
	FIN002	Income and Expenditure - Income - Tuition fees & education contracts
	FIN003	Income and Expenditure - Income - Research grants and contracts
	FIN004	Income and Expenditure - Income - Other income
	FIN005	Income and Expenditure - Income - Endowment and investment income
	FIN006	Income and Expenditure - Income - Total income
	FIN007	Income and Expenditure - Expenditure - Staff costs
	FIN008	Income and Expenditure - Expenditure - Other operating expenses
	FIN009	Income and Expenditure - Expenditure - Depreciation
	FIN010	Income and Expenditure - Expenditure - Interest and other finance costs
	FIN011	Income and Expenditure - Expenditure - Total expenditure
	FIN012	Balance Sheet - Fixed Assets - Land & Buildings
	FIN013	Balance Sheet - Fixed Assets - Equipment
	FIN014	Balance Sheet - Fixed Assets - Investments
	FIN015	Balance Sheet - Fixed Assets - Other
	FIN016	Balance Sheet - Fixed Assets - Total fixed assets
	FIN017	Balance Sheet - Debtors - Amounts falling due after one year
	FIN018	Balance Sheet - Current Assets - Fixed assets held for resale
	FIN019	Balance Sheet - Current Assets - Stocks and stores in hand
	FIN020	Balance Sheet - Current Assets - Trade debtors
	FIN021	Balance Sheet - Current Assets - Other debtors
	FIN022	Balance Sheet - Current Assets - Restricted cash and short term investments
	FIN023	Balance Sheet - Current Assets - Cash and short term investments
	FIN024	Balance Sheet - Current Assets - Total current assets
	FIN025	Balance Sheet - Creditors: Amount Falling Due Within One Year - Overdrafts
	FIN026	Balance Sheet - Creditors: Amount Falling Due Within One Year - Loans
	FIN027	Balance Sheet - Creditors: Amount Falling Due Within One Year - Capital element of finance leases
	FIN028	Balance Sheet - Creditors: Amount Falling Due Within One Year - Trade creditors
	FIN029	Balance Sheet - Creditors: Amount Falling Due Within One Year - Tax and pension contributions
	FIN030	Balance Sheet - Creditors: Amount Falling Due Within One Year - Payments on account
	FIN031	Balance Sheet - Creditors: Amount Falling Due Within One Year - Fixed asset creditors
	FIN032	Balance Sheet - Creditors: Amount Falling Due Within One Year - Other

Finance	FIN033	Balance Sheet - Creditors: Amount Falling Due Within One Year - Total current liabilities
	FIN034	Balance Sheet - Net current assets / (liabilities)
	FIN035	Balance Sheet - Total assets less current liabilities
	FIN036	Balance Sheet - Creditors: Amount Falling Due After One Year - Loans
	FIN037	Balance Sheet - Creditors: Amount Falling Due After One Year - Capital element of finance leases
	FIN038	Balance Sheet - Creditors: Amount Falling Due After One Year - Other liabilities
	FIN039	Balance Sheet - Creditors: Amount Falling Due After One Year - Total long-term liabilities
	FIN040	Balance Sheet - Provisions for liabilities
	FIN041	Balance Sheet - Net assets excluding pension asset/ (liability)
	FIN042	Balance Sheet - Net pension asset/ (liability)
	FIN043	Balance Sheet - Net assets including pension asset/ (liability)
	FIN044	Balance Sheet - Deferred capital grants
	FIN045	Balance Sheet - Revaluation reserve
	FIN046	Balance Sheet - Restricted reserves
	FIN047	Balance Sheet - I&E account excluding pension reserve
	FIN048	Balance Sheet - Pension reserve
	FIN049	Balance Sheet - I&E account including pension reserve
	FIN050	Balance Sheet - Total reserves
	FIN051	Balance Sheet - Total Funds
	FIN052	Cash Flow Statement - Net cash inflow/(outflow) from operating activities
	FIN053	Cash Flow Statement - Returns on investments and servicing of finance - Interest received
	FIN054	Cash Flow Statement - Returns on investments and servicing of finance - Interest paid
	FIN055	Cash Flow Statement - Returns on investments and servicing of finance - Interest element of finance lease rental payments
	FIN056	Cash Flow Statement - Returns on investments and servicing of finance - Net cash inflow/(outflow) from returns on investments and servicing of finance
	FIN057	Cash Flow Statement - Taxation
	FIN058	Cash Flow Statement - Capital expenditure and financial investment - Payments to acquire fixed assets
	FIN059	Cash Flow Statement - Capital expenditure and financial investment - Receipts from sale of fixed assets
	FIN060	Cash Flow Statement - Capital expenditure and financial investment - Deferred capital grants received
	FIN061	Cash Flow Statement - Capital expenditure and financial investment - Net cash inflow/(outflow) from capital expenditure
	FIN062	Cash Flow Statement - Management of liquid resources - Withdrawals or disposals (shown as positive figure)
	FIN063	Cash Flow Statement - Management of liquid resources - Deposits or acquisitions (shown as negative figure)
	FIN064	Cash Flow Statement - Management of liquid resources - Net cash inflow/(outflow) from management of liquid resources
	FIN065	Cash Flow Statement - Financing - New loans
	FIN066	Cash Flow Statement - Financing - New finance leases
	FIN067	Cash Flow Statement - Financing - Repayment of amounts borrowed - secured and unsecured loans
	FIN068	Cash Flow Statement - Financing - Repayment of finance leases (capital element)
	FIN069	Cash Flow Statement - Financing - Net cash inflow/ (outflow) from financing
	FIN070	Cash Flow Statement - Increase/ (decrease) in cash

Finance	FIN071	Cash Flow Statement - Reconciliation of net cash flow to movement in net funds/(debt) - Increase/ (decrease) in cash
	FIN072	Cash Flow Statement - Reconciliation of net cash flow to movement in net funds/(debt) - Cash to repay debt
	FIN073	Cash Flow Statement - Reconciliation of net cash flow to movement in net funds/(debt) - Cash used to increase liquid resources
	FIN074	Cash Flow Statement - Reconciliation of net cash flow to movement in net funds/(debt) - New loans and finance leases
	FIN075	Cash Flow Statement - Reconciliation of net cash flow to movement in net funds/(debt) - Change in net funds/(debt)
	FIN076	Cash Flow Statement - Reconciliation of net cash flow to movement in net funds/(debt) - Net funds/(debt) at beginning of year
	FIN077	Cash Flow Statement - Reconciliation of net cash flow to movement in net funds/(debt) - Net funds/(debt) at end of year
	FIN078	Cash Flow Statement - Analysis of net debt - Cash and short term investments
	FIN079	Cash Flow Statement - Analysis of net debt - Bank overdrafts
	FIN080	Cash Flow Statement - Analysis of net debt - Total cash and short term investments
	FIN081	Cash Flow Statement - Analysis of net debt - Debt due within 1 year
	FIN082	Cash Flow Statement - Analysis of net debt - Debt due after 1 year
	FIN083	Cash Flow Statement - Analysis of net debt - Total
	FIN084	Cash Flow Statement - Income used in ratio analysis
	FIN085	Cash Flow Statement - Ratio analysis - Liquidity - Ratio analysis - Liquidity - Cash days in hand
	FIN086	Ratio analysis - Liquidity - Adjusted cash days in hand
	FIN087	Ratio analysis - Liquidity - Current ratio
	FIN088	Ratio analysis - Liquidity - Adjusted current ratio
	FIN089	Ratio analysis - Liquidity - Trade debtors days - excluding Skills Funding Agency, EFA and HEFCE
	FIN090	Ratio analysis - Liquidity - Trade creditors days - non pay expenditure
	FIN091	Ratio analysis - Liquidity - Quick ratio
	FIN092	Ratio analysis - Cash generated from operations to income
	FIN093	Ratio analysis - Gearing - Debt charges as a percentage of income
	FIN094	Ratio analysis - Gearing - Interest as a percentage of income
	FIN095	Ratio analysis - Gearing - Total borrowing as a percentage of income
	FIN096	Ratio analysis - Gearing - Total borrowing as a percentage of reserves
	FIN097	Ratio analysis - Gearing - Total borrowing as a percentage of reserves and debt
	FIN098	Ratio analysis - Margin - Operating surplus / (deficit) after taxation (£'000)
	FIN099	Ratio analysis - Margin - Operating surplus / (deficit) as a percentage of income
	FIN100	Ratio analysis - Margin - Adjusted operating position (£'000)
	FIN101	Ratio analysis - Margin - Adjusted operating position as a percentage of income
	FIN102	Ratio analysis - Margin - Historical cost surplus/ (deficit) as a percentage of income
	FIN103	Ratio analysis - Margin - Available reserves as a percentage of income
	FIN104	Ratio analysis - Income - Dependency on European income
	FIN105	Ratio analysis - Income - Dependency on Higher Education income
	FIN106	Ratio analysis - Income - Dependency on all other income
	FIN107	Ratio analysis - Income generating activities - Contribution from other income generating activities
	FIN108	Ratio analysis - Income generating activities - Contribution from farming
	FIN109	Ratio analysis - Income generating activities - Contribution from catering, residences and conferences

Finance	FIN110	Ratio analysis - Analysis of staff costs - Staff costs as % of income (incl. contract tuition services / incl. restructuring)
	FIN111	Ratio analysis - Analysis of staff costs - Staff costs as % of income (incl. contract tuition services / excl. restructuring)
	FIN112	Ratio analysis - Analysis of staff costs - Admin. costs proportion
	FIN113	Analysis of income - Funding body grants - Higher Education Funding Council in England (HEFCE) - Higher Education
	FIN114	Analysis of income - Funding body grants - Higher Education Funding Council in England (HEFCE) - Franchised
	FIN115	Analysis of income - Funding body grants - Higher Education Funding Council in England (HEFCE) - Release of capital grants
	FIN116	Analysis of income - Funding body grants - Higher Education Funding Council in England (HEFCE) - Other HEFCE
	FIN117	Supplementary Benchmarking Information - Expenditure -Staff number and cost details and staff numbers/costs per SLN/FTE - Ave SLN learners per FTE teacher
	FIN118	Supplementary Benchmarking Information - Expenditure -Staff number and cost details and staff numbers/costs per SLN/FTE - Ave SLN learners per total non-teaching staff
Ofsted	OFS001	Ofsted Rating at time of Review
Previous QAA Review Outcome	PRV001	Outcome of previous review
	PRV003	Has ever received a negative review (1 = yes, 0 = no)
	PRV004	Outcome of previous comparable review
QAA Concerns	CON001	Count of QAA concerns raised, upheld or otherwise, since previous review
	CON002	Count of QAA concerns raised which do not relate to quality, standards, information or enhancement (and are therefore automatically invalid)
	CON003	Count of QAA concerns upheld since previous review which relate to academic standards
	CON004	Count of QAA concerns upheld since previous review which relate to the quality of learning opportunities
	CON005	Count of QAA concerns upheld since previous review which relate to the provision of information
	CON006	Count of QAA concerns not upheld since previous review which relate to academic standards
	CON007	Count of QAA concerns not upheld since previous review which relate to the quality of learning opportunities
	CON008	Count of QAA concerns upheld since previous review
	CON009	Count of QAA concerns not upheld since previous review which relate to the quality of learning opportunities

Table 6.8: The set of 181 metrics used in the HEI study prior to change-over-time and benchmarking calculations being added.

7. Predicting the Outcome of Alternative Provider Reviews

The purpose of this chapter is to determine which metrics, if any, could have predicted the outcome of past QAA reviews of alternative providers, and how accurately they could have done so. The number of comparable reviews of alternative providers undertaken means that this question can be explored both at the overall review-level, and at the more granular question-level. Moreover, in two cases it is possible for the question-level analysis to explore the prediction of specific judgements rather than the aggregated 'satisfactory' or 'unsatisfactory' outcomes.

The reason for examining whether the outcome of past QAA reviews could have been predicted is to determine whether a data-driven, risk-based approach to quality assurance could have been effective and, if so, which metrics could be useful when operating such an approach in the future. As previously discussed, however, there are significant institutional and normative challenges associated with using predictions of the likelihood of receiving one of three or four judgements for four questions to decide where to prioritise a review. Making prioritisation decisions based on up to 15 predicted probabilities per provider – the probabilities of being judged 'Meets UK expectations', 'Requires improvement to meet UK expectations' or 'Does not meet UK expectations', for four questions plus the probability of being judged 'Commended' for three of the four questions - for hundreds of providers is not practical. Furthermore, with little if any reduction in burden when only one question, rather than four, is assessed as part of a review it may not be helpful either (HEFCE, 2012b, 74). This analysis therefore begins by answering the question most likely to be of use for the future operation of a data-driven, risk-based approach in the alternative provider sector where incomplete data is not an issue:

1. Using naturally-complete metrics, could the overall outcome of QAA alternative provider reviews have been successfully predicted?

While challenging to interpret and operationalise, it may be the case that considering the likelihood of an alternative provider having quality assurance issues at a question, rather than review, level leads to great improvements in accuracy. Furthermore, considering the answer to each question at the granular 'Does not meet, 'Requires improvement to meet', and 'Meets' UK expectations level rather than the aggregated 'unsatisfactory' / 'satisfactory' level may further improve accuracy. In order to assess whether there is a significant improvement in accuracy worth accepting the additional complexity of a granular question-level approach for, the second research question focusses on the review question with the greatest number of, and best distributed, 'unsatisfactory' judgements:

2. Using naturally-complete metrics, could the exact outcome of QAA reviews of academic standards at alternative providers have been successfully predicted?

As detailed below, the result of this question is that predicting the exact outcome of the reviews, i.e. the probability a provider is likely to be 'Meets UK expectations', 'Requires improvements to meet UK expectations' and 'Does not meet UK expectations', makes no practical difference to accuracy, yet makes modelling and interpretation far more challenging. Therefore, the remaining two review questions for which there is sufficient data - *teaching and learning* and *the provision of information* – are considered to see if their outcomes could be predicted more accurately than the overall review-level outcome; however, this is done at the aggregated 'unsatisfactory' / 'satisfactory' level. The remaining questions are therefore:

3. Using naturally-complete metrics, could the aggregated outcome of QAA reviews of teaching and learning at alternative providers have been successfully predicted?
4. Using naturally-complete metrics, could the aggregated outcome of QAA reviews of the provision of information at alternative providers have been successfully predicted?

Standardisation of the data by year has not been considered for alternative providers as the reviews have all taken place within the space of two years. The chapter begins with an overview of the alternative provider sector and its unique challenges. This is followed by a step-by-step description of the analysis and results, and finally a brief discussion of the findings.

7.1. Introduction

The term 'alternative provider' is confusing and something of a catch-all. BIS defines 'alternative providers' (in England-only presumably) as "higher education providers who do not receive funding from, and are not regulated by, the Higher Education Funding Council for England (HEFCE)" (BIS, 2015a, 1), but such a definition could be expanded to FECs. In reality, an 'alternative' provider is a provider that is not a Higher Education Institution (HEI) as designated by the 1992 Higher Education Act and is not an FEC. Whilst all alternative providers are private, not all private providers are alternative providers, the University of Buckingham for example is a private HEI.

To make matters more confusing, not all alternative providers in England appear on HEFCE's "Register of Higher Education Providers" (HEFCE, 2015b). This is because only providers which meet one of the below criteria are featured:

- receive direct public grants for HE;

- have courses which have been specifically designated by Government as eligible for the purposes of English student support funding;
- are higher education institutions (HEIs);
- and / or have the right to award one or more types of UK degree.

Therefore, those alternative providers that do not provide designated courses but have either unsuccessfully applied for course designation or have applied for HTS, successfully or otherwise, will have been reviewed by the QAA but will not feature on HEFCE's register. The Bedfordian Business School, for example, has applied for HTS and teaches Pearson HNCs and HNDs as well as programmes franchised from the University of Maribor in Slovenia (QAA, 2014e) but no UK-based degrees; thus it was reviewed by the QAA but does not feature on HEFCE's register. No register of higher education providers exists for Scotland, Wales or Northern Ireland; however, the vast majority of alternative providers are based in England – 264 out of the 273 contained in the data set.

The lack of clarity over what constitutes an alternative provider makes determining the size of the sector difficult. As of August 2015 HEFCE's register contained 117 alternative providers which have specific courses that are eligible under student support regulations and the QAA had records of 428 alternative providers (HEFCE, 2015c). There are an unknown number of providers offering international awards that have not sought highly trusted status or course designation and therefore do not fall under the remit of the QAA or national funding councils.

What is clear is that the number of alternative providers, and the number of students studying at them, has grown dramatically since the 2011 White Paper sought to "make it easier for new providers to enter the sector" by removing "the regulatory barriers that are preventing a level playing field for higher education providers of all types, including further education colleges and other alternative providers" (BIS, 2011, 10). Just two academic years after the White Paper had been released, the number of students studying with alternative providers grew from 7,000 to 53,000 with half of the total growth accounted for by just five providers (NAO, 2014). Data from the Student Loans Company shows that the amount paid out in tuition fee loans to full-time students at alternative providers in England-only rose from £16M in 2010/11 to £192M in 2013/14 (author analysis based upon SLC, 2015).

The hundreds of organisations that make up the alternative provider sector in the UK are diverse. Founded in 1992, granted degree-awarding powers in 2007 and the 'university' title in 2013, BPP University delivers 'education of the professions', such as business and law, across 12 locations in the UK. Conversely, in 2013/14, there were a minimum of 32 providers, many of which were very

young, with 50 students or fewer[8]. Indeed, the median age of an alternative provider in the data set used for this study, based upon their date of incorporation, was under 9.5 years. The overwhelming majority of provision is focused on business and management; however, there are a number of niche providers focusing on the arts, faith, and even needlework.

When loans for students on alternative providers' designated courses where introduced in 2011/12, no limits were placed on the number of students that providers could recruit and some expanded rapidly. By the time the Government introduced controls in November 2013, having spent far more than it had budgeted for, Regent's College had expanded from having 10 HND students to over 1,000. Similarly, the intake at St. Patrick's College ballooned from 50 to over 4,000 in one year (McGettigan, 2014). Concerns over rapid expansion were accompanied by evidence of very high drop-out and absence rates, poor administration and inappropriate recruitment practices including colleges recruiting on the streets and students being accepted onto courses while lacking adequate English language skills (PAC, 2015). The QAA's review activity shows that these concerns were to some extent justified. 42 of the 328 alternative provider reviews considered in this study, 12.8%, resulted in an 'unsatisfactory' overall judgement. This is higher than the 7.8% for FECs and 6.3% for universities; however, 286 out of 328 reviews resulting in a 'satisfactory' judgement overall still represents a broadly compliant sector.

7.2. Results – Review-Level Outcomes

Using naturally-complete metrics, could the overall outcome of QAA alternative provider reviews have been successfully predicted?

7.2.1. Initial Data Exploration

The first step in the analysis was to explore which individual metrics had a strong relationship with the overall 'satisfactory' or 'unsatisfactory' outcome of the reviews. The only metrics with a p-value less than 0.25 were financial metrics developed from the providers' financial accounts

[8] The figure of at least 33 alternative providers with 50 or fewer students was derived from Student Loan Company Figures. These figures show that 1,614 full-time students took out student loans to cover tuition fees for courses at alternative providers that had 50 students or fewer in total. The number of such providers was not detailed and therefore 33 represents the absolute minimum, i.e. it assumes each provider had 50 students. In reality the number of such alternative providers is likely much greater.

and the provider's age, defined as the number of days since their incorporation, at the time of their review:

Metric Code	Metric Description	P-value
APA001	Age at time of review	0.003
APA011	Cash at Bank and in Hand	0.017
APA018	Total Net Assets/ (Liabilities)	0.034
APA006	Tangible Assets	0.057
APA013	Creditors: Amounts Falling Due Within One Year	0.066
APA003	Satisfied Mortgage Charges	0.247

Table 7.1: A breakdown of all metrics from the overall review level alternative provider data set with a p-value of less than 0.25.

Unlike some of the metrics with low p-values for HEIs and FECs, one can conceive these finance metrics having an impact on quality assurance processes: providers who are troubled financially may be less focused on quality and the processes that assure it than those who are financially sound.

Figure 7.1 below shows the metric with the smallest p-value, *APA001 – the provider's age at the time of their review in days.* None of the older providers have been found 'unsatisfactory':

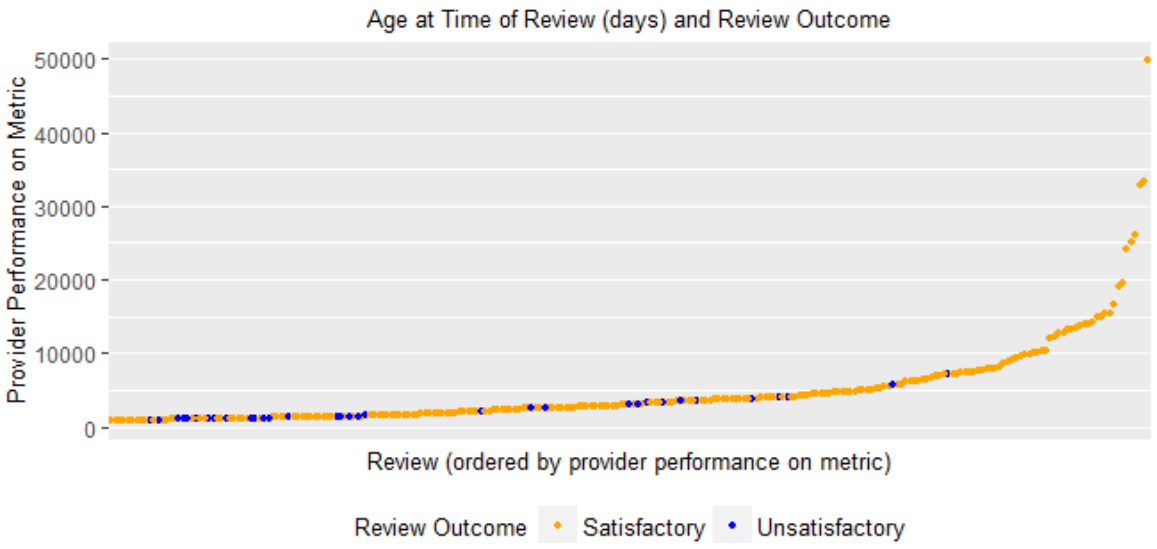

Figure 7.1: A plot of the provider's age at the time of their review and the overall outcome of that review.

There also appears to be a clustering of younger providers that have been judged 'unsatisfactory'; however, this appears to be an effect of scaling and when the much older providers are removed from the plot the clustering is not so apparent:

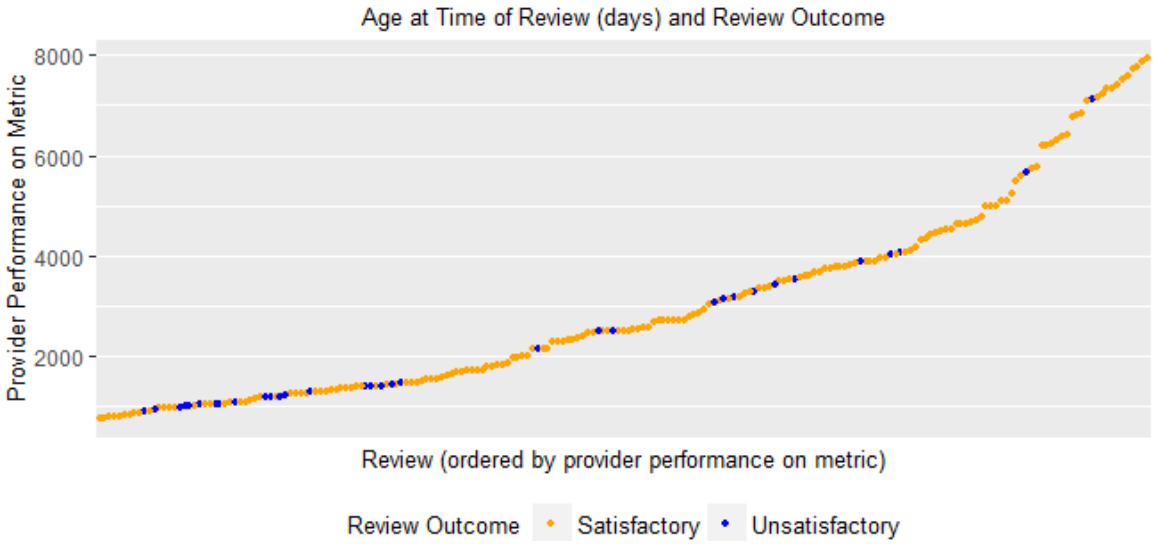

Figure 7.2: A plot of the provider's age at the time of their review and the overall outcome of that review for providers established less than 8,000 (c.21years 11 months) days at the time of their review.

This again appears to be a case of the metric being useful in identifying providers that are likely to be 'satisfactory', but less useful in identifying providers likely to be 'unsatisfactory' which are spread out amongst a large number of 'satisfactory' providers.

The same pattern is present for the amount of money providers have 'at the bank or in hand'. Figure 7.3 below shows that those providers with a relatively large amount of cash at the bank or in hand have always been found 'satisfactory':

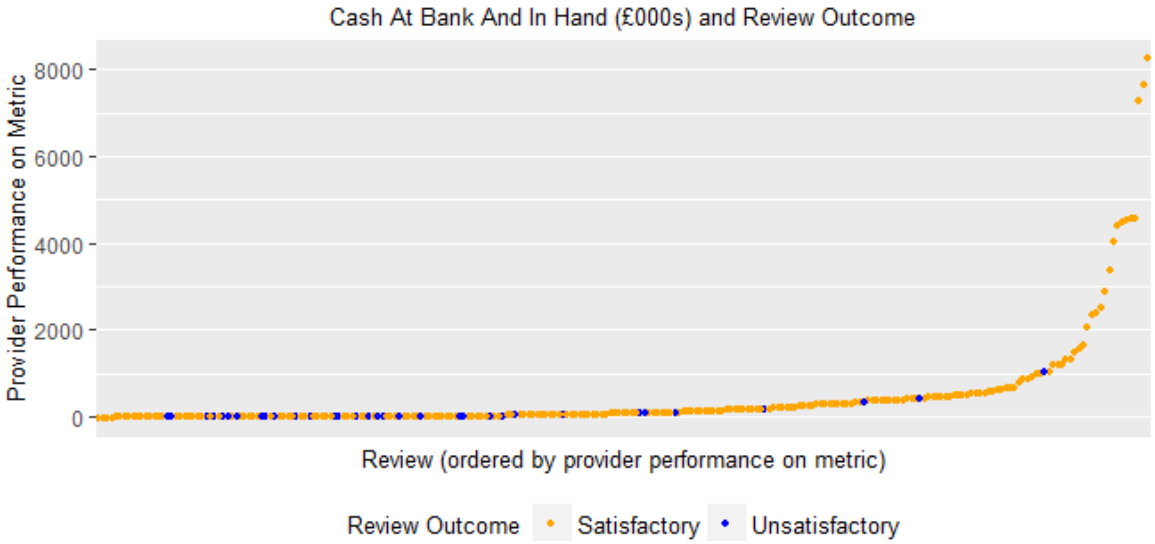

Figure 7.3: A plot of each provider's cash at the bank or in hand according to their latest set of accounts prior to their review and the overall outcome of that review.

However, focusing on the providers with less than £1,100,000 in cash at the bank or in hand shows that there are a number of 'satisfactory' and 'unsatisfactory' providers with no 'cash at the bank or in hand'. Moreover, there are also a number of 'unsatisfactory' providers with several hundred thousand pounds 'at the bank or in hand'.

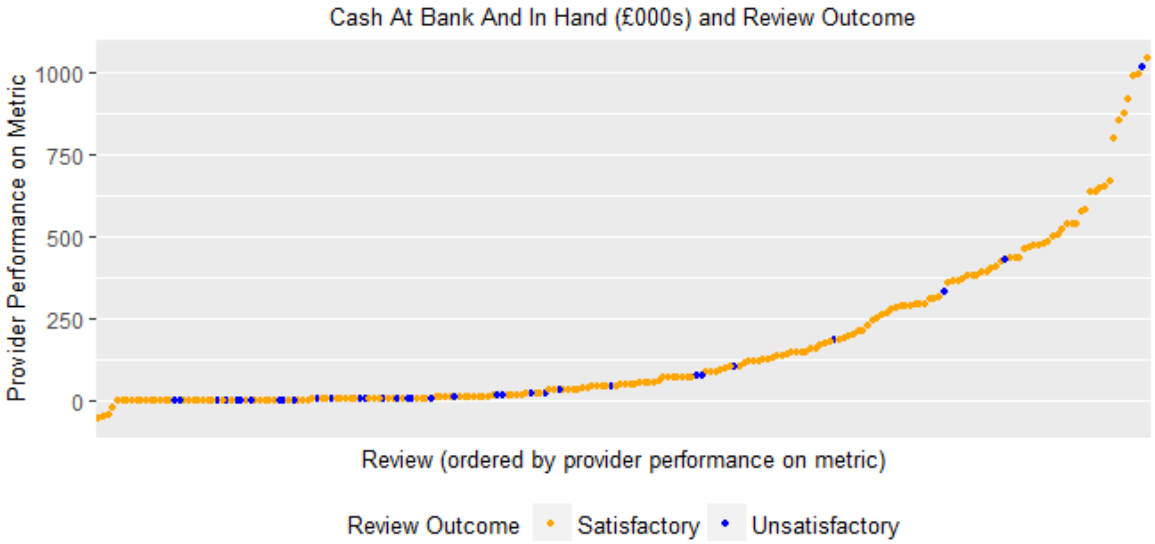

Figure 7.4: A plot of each provider's cash at the bank or in hand according to their latest set of accounts prior to their review and the overall outcome of that review for providers with less than £1.1M cash.

Again, the pattern is repeated when we examine each provider's total net assets or liabilities: there is a group of providers with relatively substantial assets who have all been found 'satisfactory', and a group of providers with no assets or liabilities most of whom have been found 'satisfactory' interspersed with the 'unsatisfactory' providers. Interestingly, all of the providers with greater liabilities than assets, indicated by points below 0 on the y-axis, have been found 'satisfactory':

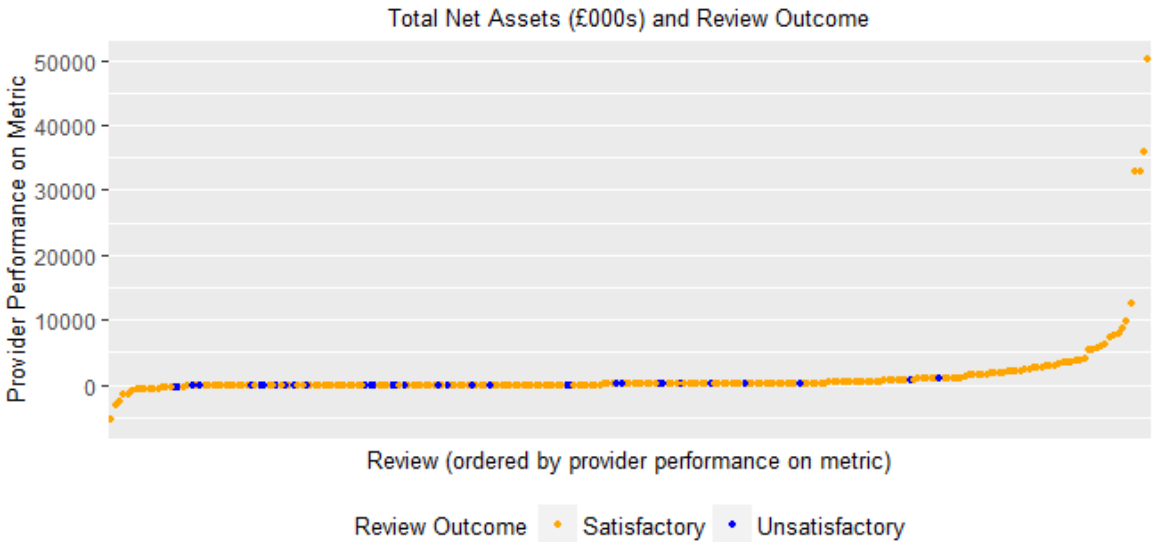

Figure 7.5: A plot of the provider's total net assets/(liabilities) at the time of their review and the outcome of that review.

Examining the outcome of the previous comparable reviews shown below in figure 7.6 reveals that the majority of reviews were the first that the provider had undergone:

Figure 7.6: The outcome of provider's reviews following their previous comparable review broken down by the outcome of the previous comparable review.

All seven reviews of providers that had already had a negative review were 'satisfactory'. Whilst one could argue this is a testament to the review process with all providers putting right their areas for concern, the sample size is too small to draw any meaningful conclusion. An alarming five out of the 46 reviews which followed on from a previous positive review found the provider to be 'unsatisfactory' showing a worrying ability for performance to decline.

In summary, there are a small number of financial metrics which allow us to identify which providers are most likely to be 'satisfactory', but are of limited use when trying to distinguish the 'unsatisfactory' providers. 'Satisfactory' performance in a previous comparable review is, if anything, an indication of an increased likelihood of being 'unsatisfactory' according to the data.

7.2.2. Fitting the Model

Running the *elastic net* procedure results in the diagnostic plots shown below in Figure 7.7:

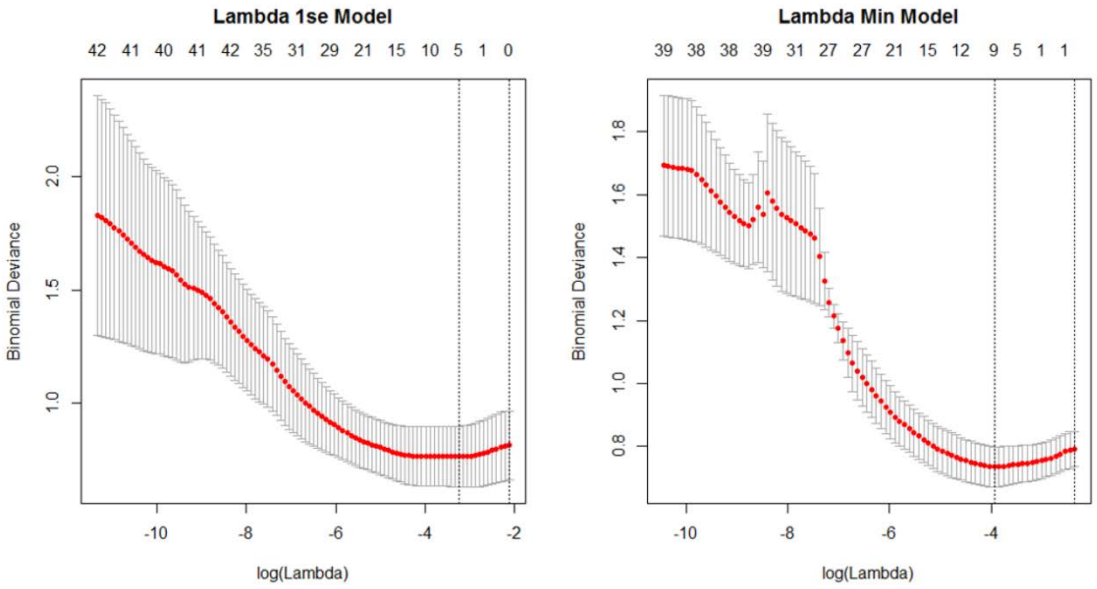

Figure 7.7: The diagnostic plots for the λ_{1se} (left) and λ_{min} (right) model for alternative provider overall review level outcomes.

On this occasion the choice between the λ_{1se} model and the λ_{min} model is not straightforward. The ostensibly more accurate λ_{min} model contians seven metrics:

- APA004 - Financial Accounts Type (e.g. 'Total exemption', 'Full').
- PRV002 - Has the provider been reviewed and received a negative outcome in the last review?
- PRV003 - Has ever received a negative review?
- APA001 - Age at time of review.
- APA011 - Amount of Cash At Bank And In Hand (£000's).
- CON002 - Count of *QAA concerns* raised which do not relate to quality, standards, information or enhancement (and are therefore automatically invalid).
- CON005 - Count of *QAA concerns* upheld since previous review which relate to the provision of information.

Whereas the λ_{1se} model, which represents a simpler but less accurate model within specified limits contains a subset of four of those metrics:

- APA004 – Financial Accounts Type (e.g. 'Total exemption', 'Full').
- PRV002 - Has the provider been reviewed and received a negative outcome in the last review?
- PRV003 - Has ever received a negative review?
- APA001 - Age at time of review.

The specific nature of the CON005 metric and the very small relative size of the coefficients for the additional metrics included in the λ_{min} model suggest that these additional metrics only serve to overfit the model, or at the very least complicate it for very minimal gain. For that reason the λ_{1se} model is explored further.

The specific λ_{1se} model calculates the probability of an unsatisfactory review as:

$$\frac{e^A}{1+e^A}$$

where:

$$A = -2.25 + (0.84 \times APA004.TES) + (-0.25 \times APA004.FUL) + (-0.12 \times PRV002.YES + (-0.10 \times PRV003.YES) + (-0.00002 \times APA001)$$

and:

APA004.TES = 1 if the provider is classified as a 'Total exemption SMALL' body by Companies House, 0 otherwise
APA004.FUL = 1 if the provider is classified as a 'FULL' body by Companies House, 0 otherwise
PRV002.YES = 1 if the provider been reviewed and received a negative outcome in the last review, 0 otherwise
PRV003.YES =1 if the provider has ever received a negative review, 0 otherwise

As the coefficients of the APA004.FUL, PRV002.YES and PRV003.YES metrics are negative, the result of these conditions being met, i.e. the provider is classified as a 'FULL' body by Companies House, the provider having previously received a negative outcome in their last review, and if the provider has ever received a negative review, is that the likelihood of the provider being 'unsatisfactory' decreases. Likewise the older the provider, i.e. the larger the value of APA001, the lower the predicted likelihood of the provider being 'unsatisfactory'. With a positive coefficient, a provider classed as 'Total exemption SMALL' by Companies House will have an increased predicted likelihood of being judged 'unsatisfactory'.

Larger providers being less likely to be 'unsatisfactory' is perhaps unsurprising, the larger the organisation the more effort and resource can be dedicated to quality assurance and quality assurance reviews. Those providers that have previously been found 'unsatisfactory' subsequently being less likely to be 'unsatisfactory' than those providers that haven't is somewhat counterintuitive but can be understood. A provider that has previously been judged 'unsatisfactory' will have an understanding of the areas where it needs to improve and will be motivated to work hard to avoid a reoccurence.

7.2.3. Evaluating the Model

Having given worked examples of how each model works in the previous chapters a further demonstration will not be given for the alternative provider models. The next stage is therefore to evaluate how well the model fits the data with which it was developed.

7.2.3.1. Testing the Fit of the Model

Figure 7.8 below shows the ROC curve for this model when applied to the data used to develop it. The 'area under the curve' value of 0.778 suggests a reasonable rate of 'unsatisfactory' FECs being succesfully prioritised as the threshold criteria for triggering a review is lowered:

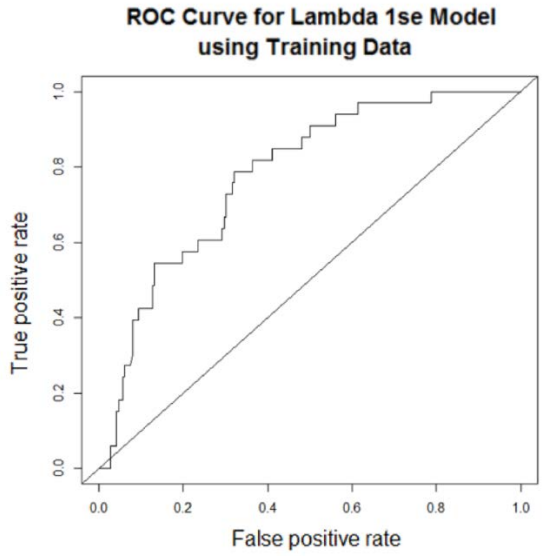

Figure 7.8: The ROC curve for the λ_{1se} model fitted to the training data set for overall review level, alternative provider outcomes.

Figure 7.9 and Table 7.2 below show the effect of lowering the threshold required for the model's predicted probability of failure to trigger a review.

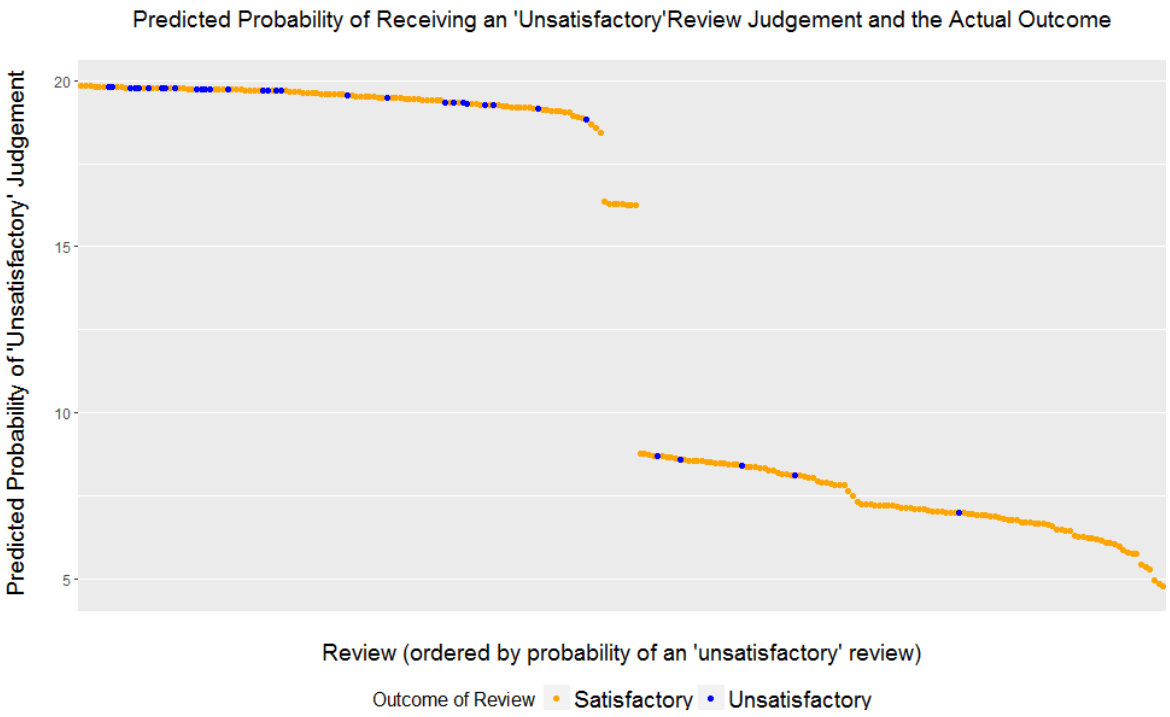

Figure 7.9: Predicted probabilities for each of the 245 complete, comparable reviews used to train the model and their actual outcome.

	Predicted Probability of an 'unsatisfactory' outcome required to trigger a review	Number of 'unsatisfactory' reviews (true positives)	Number of 'satisfactory' reviews (false positives)	Accuracy rate
← Decreasing probability of a review being 'unsatisfactory' required to trigger inspection	19.46%	5	9	0.84
	19.43%	10	17	0.82
	19.38%	15	27	0.79
	19.12%	20	50	0.74
	18.90%	25	67	0.70
	9.24%	30	106	0.64
	7.21%	33	167	0.55

Table 7.2: The number of 'satisfactory' and 'unsatisfactory' alternative provider reviews that would have resulted from decreasing the threshold required to prompt a review (only select points are shown).

The model performs well to begin with, with the first nine 'unsatisfactory' reviews being predicted with just 13 'satisfactory' reviews being prioritised. However, as with previous models the error rate then begins to increase and 167 'satisfactory' reviews would have been incorrectly prioritised before all of the 'unsatisfactory' reviews had been conducted. This, with perfect hindsight, is an error rate of 167 / (33 + 167) = 83.5% amongst the providers prioritised for review; worse than that for FECs and again unlikely to sit well with the alternative providers being priotised under the system.

The high error rate is once again a result of the narrow range of predicted proabilities: the relationship between the metrics and the outcome of the reviews is weak resulting in uncertain predictions. Althougth there is more differentiation between the predicited probabilities compared to the HEI and FEC models, some of this is simply due to three of the four metrics in the model being discrete. The actual predicted probabilities still differ little with no review forecast as being 'unsatisfactory' with a probability of greater than 20%.

The model therefore appears to be of little benefit if the policy goal is to identify all the 'unsatisfactory' performance. Even with the unrealistic benefit of perfect hindsight allowing the QAA to stop performing reviews imediately after the final 'unsatisfactory' provider has been reviewed, 200 out of 245 reviews would have been required. If the policy goal is to identify a high-risk group of providers of which a greater proportion turn out to be 'unsatisfactory' compared to the low-risk group, acccepting that in a risk-based system some failure must be tolerated, then this appears possible. Abritrarily dividing the providers into equally-sized low and high risk groups – i.e. the riskiest 50% being categorised as 'high risk' and the least risky 50% being categorised as 'low risk' – 28 out of the 33 'unsatisfactory' providers would be in the high-risk group.

7.2.3.2. Assessing the Model's Predictions

As discussed previously in the methods chapter, the number of reviews of alternative providers allowed for 20% of the data to be held back when developing the model specifically to test its performance. Figure 7.10 below show the ROC curve for the model's application to the training set comprising 60 reviews:

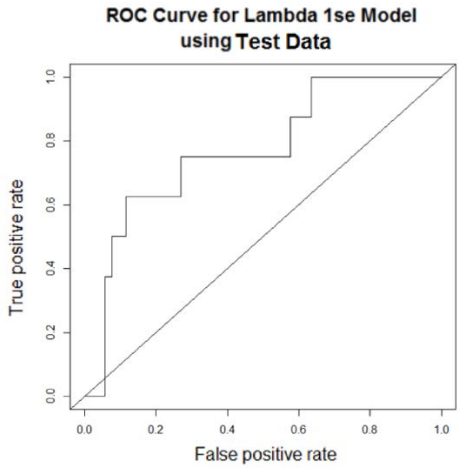

Figure 7.10: The ROC curve for the λ_{1se} model fitted to the test data set for overall review level outcomes.

The ROC Curve has a similar appearance and similar AUC value of 0.769 to that of the model. Figure 7.11 below shows the effect of lowering the threshold required to trigger a review for the test data.

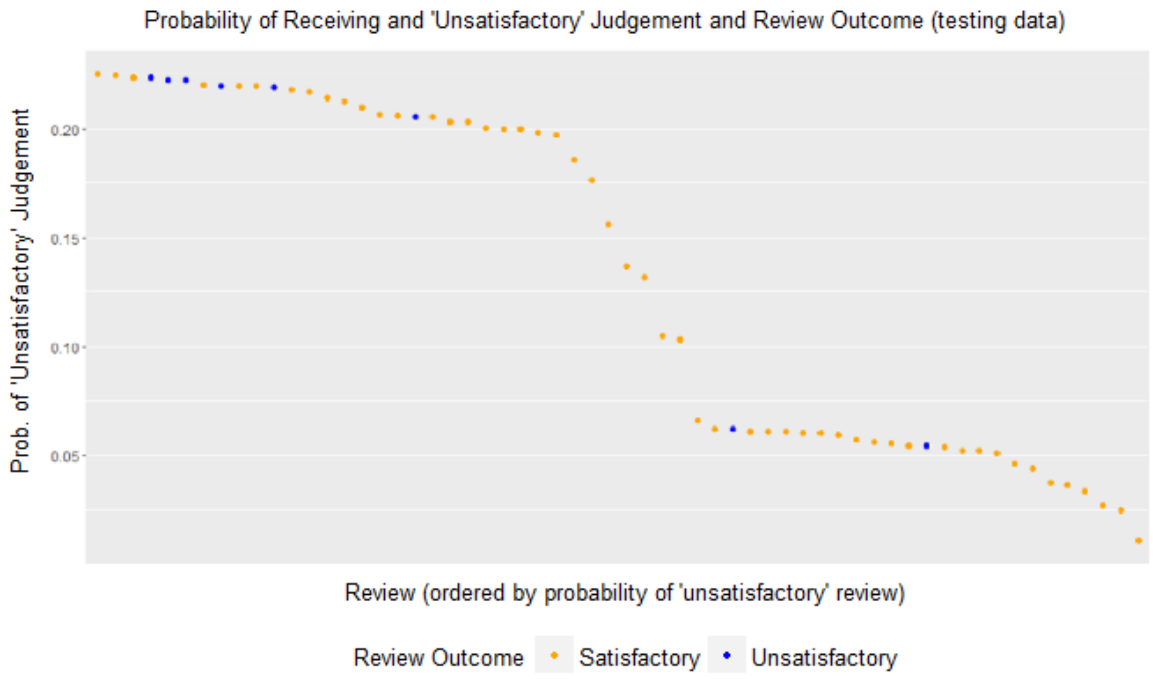

Figure 7.11: The predicted probabilities and actual outcomes of the reviews contained in the testing set for alternative providers.

When applied to new data the model reassuringly performs similarly to when it is applied to the training data and better than other models have performed previously. Of the first eight reviews that would have been prioritised, four would have resulted in an 'unsatisfactory' judgement; however, as is a familiar pattern now, after some initial success there are a number of providers that are not prioritised until after a significant number of 'satisfactory' reviews. Even if the QAA had the ability to stop reviewing alternative providers at the optimal point immediately after the least likely 'unsatisfactory' review has been conducted, the model would still have had an error rate amongst reviewed providers of 40 / (40 + 8) = 83.3%. The model still performs well if the policy goal is merely to identify a high and low risk group: 6 out of 8 'unsatisfactory' providers appear in the riskiest half of providers.

The model's similar performance when applied to the training and test data reassures us that the model is not overfit, it has not learned the specific 'noise' arising from the training set and is picking up on genuine (albeit weak) patterns in the data, i.e. that smaller providers and those who have not previously been judged 'unsatisfactory' are slightly more likely to be so in the future. This is not universally the case however.

As discussed in the earlier methods chapter there are two tests that we have not been able to perform due to the resource demands of acquiring new data. The first is to apply the model to recent reviews to see if the relationship still holds. The FEC analysis showed that relationships which had previously existed no longer do so making the model operationally useless. The second is to pick a specific point in time and calculate each provider's risk score. If a large number of those not reviewed within a year were predicted as more likely to be 'unsatisfactory' than those 'unsatisfactory' providers that were reviewed, then either there is a lot of 'unsatisfactory' provision going unreviewed or the model may not work so well in real time. Any perceived success must therefore be viewed with caution.

7.2.3.3. Summary

The best model calculated the predicted probability of an alternative provider being judged 'unsatisfactory' overall based on previous review performance and the age and size of the provider. The model determined the predicted likelihood of being judged 'unsatisfactory' as:

$$\frac{e^{A}}{1+e^{A}}$$

Where:

$$A = -2.25 + (0.84 \times APA004.TES) + (-0.25 \times APA004.FUL) + (-0.12 \times PRV002.YES + (-0.10 \times PRV003.YES) + (-0.00002 \times APA001)$$

And:

APA004.TES = 1 if the provider is classified as a 'Total exemption SMALL' body by Companies House, 0 otherwise
APA004.FUL = 1 if the provider is classified as a 'FULL' body by Companies House, 0 otherwise
PRV002.YES = 1 if the provider been reviewed and received a negative outcome in the last review, 0 otherwise
PRV003.YES =1 if the provider has ever received a negative review, 0 otherwise

In some regards the model was the most promising yet: the predictions based on the held-back testing data reassures us that the model was picking up on weak but genuine underlying patterns in the data and the metrics themselves make intuitive sense. One can conceive how larger, more established providers, and those that have previously been found 'unsatisfactory' and been required to focus on improving could be less likely to be found 'unsatisfactory'. The model still faces significant challenges in that, as with all previous models, a significant number of 'satisfactory' reviews would have to be prioritised in order to successfully identify all the 'unsatisfactory' providers. There is a large amount of uncertainty in the model and an error rate that the sector, if not the QAA also, would likely find unacceptable. The model did however do a

reasonable job of identifying a 'high risk' group which contained the majority of 'unsatisfactory' providers– 6 out of 8 in the test dataset. Therefore, the answer to the question

> *Using naturally-complete metrics, could the overall outcome of QAA alternative provider reviews have been successfully predicted?*

is dependent on the policy definition of success. If any 'unsatisfactory' provision going unreviewed is unacceptable then the answer is no, near enough all providers would have needed to have been reviewed in order for all 'unsatisfactory' provision to have been identified. If some 'unsatisfactory' provision going unreviewed is acceptable, then the model would have managed to have prioritised the majority of 'unsatisfactory' providers in the first 50% of providers to be reviewed. Even then, the weak underlying relationships that inform the model are susceptible to breaking down completely in a rapidly evolving sector meaning there is no guarantee the model would continue to work in the future.

7.3. Results – Academic Standards

Using naturally-complete metrics, could the exact outcome of QAA reviews of academic standards at alternative providers have been successfully predicted?

Putting aside the normative challenges of combining different probabilities for the different questions that make up a QAA review, the ability to successfully predict the more granular 'Does not meet UK expectations', 'Requires improvement to meet UK expectations' or 'Meets UK expectations' outcome for each review question represents the ideal for a fully risk-based approach to prioritisation. If the QAA could predict the outcome, it would then be able to prioritise those providers most likely to be judged 'Does not meet UK expectations' first, before the providers most likely to be judged 'Requires improvement to meet UK expectations' next, in a way that focusing on the provider deemed most likely to be 'unsatisfactory' may not allow. As there are sufficient numbers of each ordinal judgement ('Does not meet UK expectations', 'Requires improvement to meet UK expectations', and 'Meets UK expectations') for the *academic standards* question, this analysis will explore whether it is possible, and helpful, to predict more granular review outcomes.

7.3.1. Initial Data Exploration

The first step in the analysis was to examine how similar the outcome of the *academic standards* question was to the overall outcome of reviews of alternative providers. Table 7.3 below shows

that of 41 reviews which were 'unsatisfactory' overall, 31 were 'unsatisfactory' in relation to *academic standards.*

	Unsatisfactory		Satisfactory
	Does not meet UK expectations	Requires improvement to meet UK expectations	Meets UK expectations
Overall review outcome	41		264
Academic standards outcome	11	20	270

Table 7.3: The results for the overall outcome and academic standards section of each alternative provider review. Note that four reviews did not assess academic standards.

With over three quarters of providers judged 'unsatisfactory' overall having also been judged 'unsatisfactory', i.e. either 'Does not meet UK expectations' or 'Requires improvement to meet UK expectations', in relation to *academic standards* it is likely the models for predicting the outcome of the overall review and the outcome of the *academic standards* question will be similar.

The second step was to explore which individual metrics had a strong relationship with the ordered 'Does not meet UK expectations', 'Requires improvement to meet UK expectations' and 'Meets UK expectations' outcomes of the *academic standards* section of the reviews. Seven metrics had a p-value less than 0.25: four finance metrics, three of which were also significant in relation to the overall review findings in the previous section, two *QAA Concerns* metrics, and the provider's age at the time of their review.

Metric Code	Metric Description	P-value
APA009	Investments / Stocks	0.000
APA011	Cash at Bank and In Hand	0.009
APA001	Age at time of review	0.010
CON003	Count of QAA concerns upheld since previous review which relate to academic standards	0.087
APA006	Tangible Assets	0.105
APA013	Creditors: Amounts Falling Due Within One Year	0.123
CON005	Count of QAA concerns upheld since previous review which relate to the provision of information	0.156

Table 7.4: A breakdown of all metrics from the academic standards alternative provider data set with a p-value of less than 0.25.

It is easy to conceive why the *count of QAA concerns upheld since previous review which relate to academic standards* metric has a significant relationship with the outcome of the review of alternative providers' academic standards: those providers for which the QAA have investigated concerns and found 'unsatisfactory' performance are likely to have issues which may not be resolved by the inevitable, soon-to-follow review. The nature of the relationship between review outcomes relating to *academic standards* and the *CON005 - QAA concerns relating to the provision of information* metric is less apparent.

Figure 7.12 below shows the amount of money providers have 'at the bank or in hand'. As was the case for the overall model in the previous section, those providers with a relatively large amount of cash at the bank or in hand have always been found 'satisfactory'. What also becomes clear here is that there is no differentiation between the 'cash at bank and in hand' of providers judged 'Does not meet UK expectations' and providers judged 'Requires improvement to meet UK expectations'. This again appears to be a case of the metric being useful in identifying providers who are likely to be 'satisfactory', but less useful in identifying providers likely to be 'unsatisfactory' – regardless of the degree to which they are 'unsatisfactory - which are spread out amongst a large number of 'satisfactory' providers.

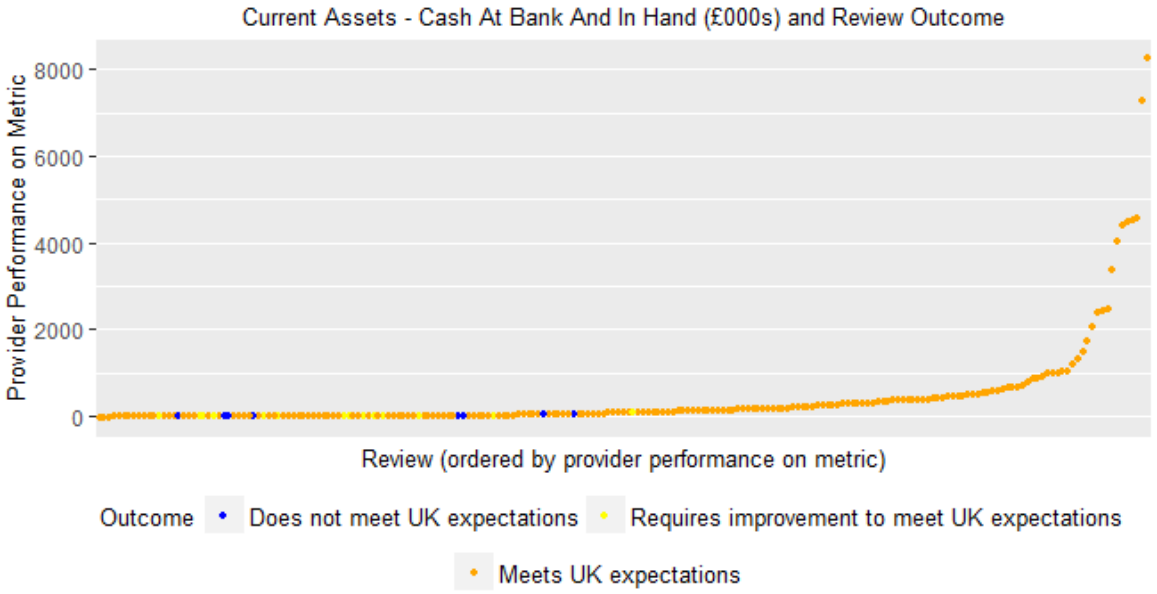

Figure 7.12: A plot of each provider's cash at the bank or in hand according to their latest set of accounts prior to their review and the outcome of the academic standards question of that review.

As with the analysis for the overall review outcome, focusing on the providers with less than £1,100,000 in 'cash at the bank or in hand' shows the pattern more clearly. There is no difference in metric performance for those judged 'Requires improvement to meet UK expectations' and those judged 'Does not meet UK expectations':

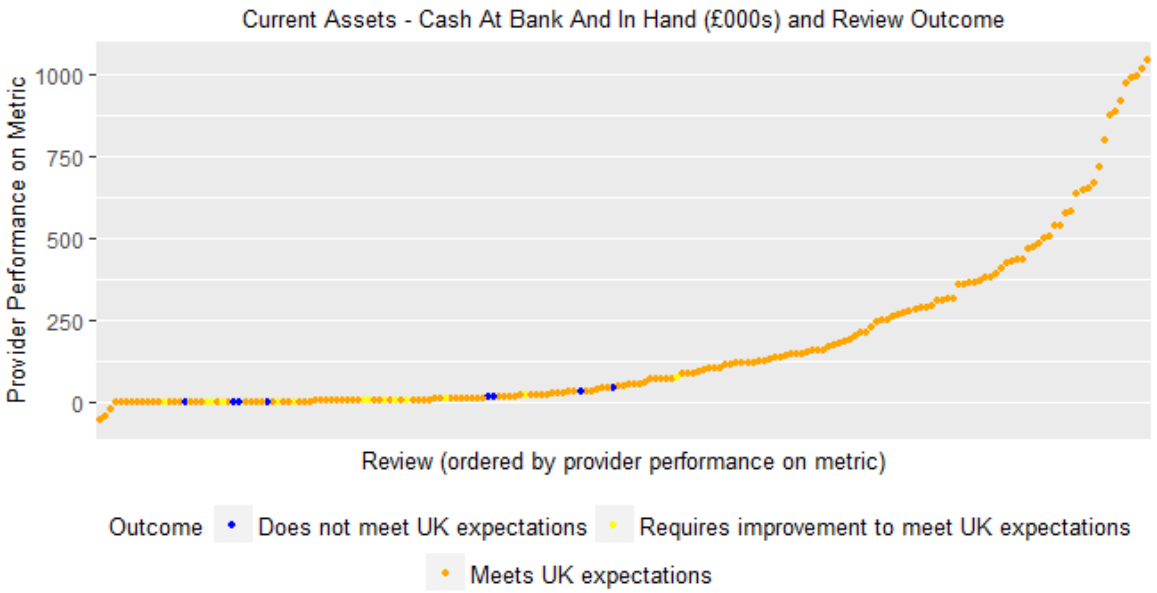

Figure 7.13: A plot of each provider's cash at the bank or in hand according to their latest set of accounts prior to their review and the outcome of the academic standards question of that review for providers with less than £1.1M cash.

On closer inspection the *CON003 - count of QAA concerns upheld since previous review which relate to academic standards* metric seems somewhat affected by low numbers. Only two alternative providers have had *QAA concerns* relating to academic standards upheld, one provider which had two such concerns was subsequently found to 'Require improvement to meet UK expectations' whereas the second which had had a single *QAA concern* upheld was found to 'Meet UK expectations' on the subsequent review. The same is also true for the *CON005 - count of concerns upheld since previous review which relate to the provision of information* metric as both the providers which had *QAA concerns* relating to academic standards upheld also had *QAA concerns* relating to the provision of information upheld at the same time.

In summary, the majority of alternative providers judged 'unsatisfactory' overall were judged 'unsatisfactory' in relation to *academic standards*. One would therefore expect to see similarities in the significant metrics for the overall and *academic standards* models. These similarities were present with the small number of financial metrics which can identify which providers are most likely to be judged 'Meets UK expectations'. Also significant for the *academic standards* model were two *QAA Concerns* metrics; however, on closer inspection this is an anomaly as the result of the rarity of *QAA concerns* being upheld. Again the metrics appear to be able to identify a subset of alternative providers most likely to be performing to the desired standard but are less able to identify those most likely to be performing below this standard.

7.3.2. Fitting the Model

One part of the process that remains unchanged whether the dependent variable is the binary 'satisfactory' / 'unsatisfactory' outcome or the more granular judgements is selecting the best model.

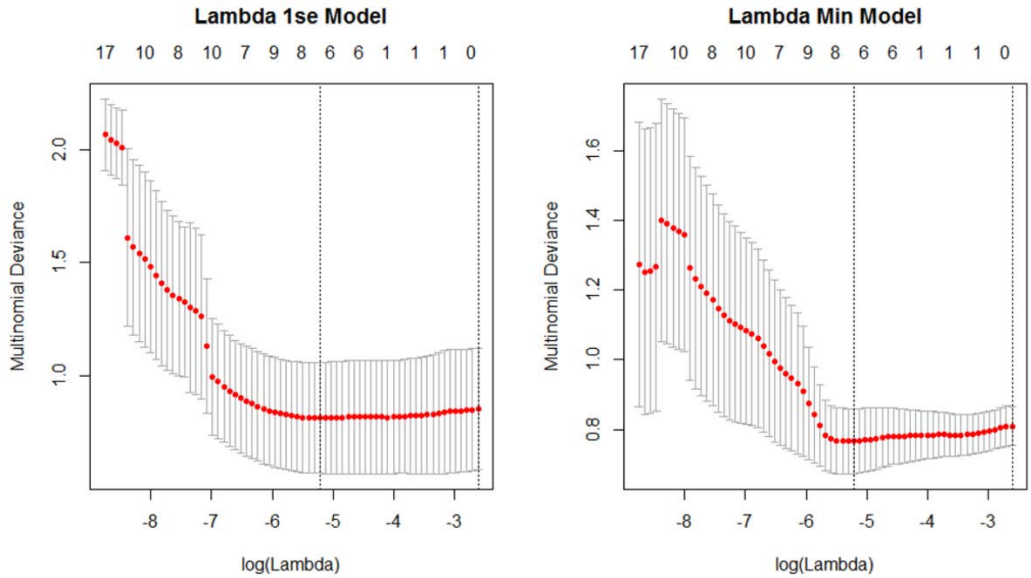

Figure 7.14: The diagnostic plots for the λ_{1se} (left) and λ_{min} (right) model for the multinomial academic stadnards model.

On this occasion there is little difference between the λ_{1se} model (lefthand plot) and the λ_{min} model (righthand plot). At the point where they are at their most accurate, both models include a small number of variables with very small coefficients suggesting a degree of overfitting. At the point where each model is at its most parsimonious – i.e. it has the fewest metrics with a fit that is within one standard deviation of the minimum – it contains no metrics whatsoever. That result means that the QAA could be best off by ignoring the data entirely and assuming each provider had the same chance of being of being judged 'Does not meet UK expectations' (11/301 = 3.7%), 'Requires improvement to meet UK expectations' (20/301 = 6.6%), or 'Meets UK expectations' (270/301 = 89.7%). The best fitting and most appropriate model is the λ_{min} model considered further below.

One aspect of the analysis that is different when there are more than two possible review outcomes is how the model is constructed. In effect three models are created, each to predict the likelihood of each of one of the three outcomes occuring and comprising an independent set of metrics. The probability for each provider being judged 'Does not meet UK expectations', 'Requires improvement to meet UK expectations' and 'Meets UK expectations' will always total one. The three models which constitute the λ_{min} model contain the following metrics:

To predict the likelihood of being judged 'Does not meet UK expectations':

- APA004 – Financial Accounts Type (e.g. 'Total exemption', 'Full').
- PRV001 - Outcome of previous review
- PRV004 - Outcome of previous comparable review
- PRV006 - Worst judgement in previous comparable review
- PRV009 - Worst judgement concerning learning in previous comparable review
- PRV010 - Outcome concerning learning in previous comparable review
- PRV012 - Outcome concerning information in previous comparable review

To predict the likelihood of being judged 'Requires improvement to meet UK expectations':

- APA004 – Financial Accounts Type (e.g. 'Total exemption', 'Full').
- PRV001 - Outcome of previous review
- APA002 – Count of the number of outstanding mortgage charges against the provider[9].
- APA014_Ca1i – The change in Net Current Assets/ (Liabilities) since the last set of accounts was filed.
- CON003 - Count of *QAA concerns* upheld since previous review which relate to academic standards

To predict the likelihood of being judged 'meets UK expectations':

- APA004 – Financial Accounts Type (e.g. 'Total exemption', 'Full').
- PRV002 - Has the provider been reviewed and received a negative outcome in the last review?
- PRV003 - Has ever received a negative review?
- APA001 - Age at time of review.
- APA009 – Amount of Investments / Stocks (£000's) held.
- APA011 – Amount of Cash At Bank And In Hand (£000's).
- APA016 - Creditors: Amounts Falling Due After One Year (£000's).
- CON002 - Count of *QAA concerns* raised which do not relate to quality, standards, information or enhancement (and are therefore automatically invalid).

There is significant overlap in the metrics used in each model. In each case the specific probabilities are given by:

[9] A mortgage charge is the means by which lenders enforce their rights to a property in the event of a debt not being paid. The greater the number of outstanding mortgage charges, the greater the number of current debtors the company has.

$$P(Does\ not\ meet\ UK\ expectations) = \frac{e^A}{1+e^A}$$

Where, to one significant figure:

$$A = \big(-1.8 + (4.6 \times APA004.TES) - (0.2 \times PRV001.POS)\big) + (0.7 \times PRV004.NPC) + (4.E-14 \times PRV006.NPR) + (1.E-15 \times PRV009.NPR) + (4.E-15 \times PRV010.NPR) + (1.E-16 \times PRV012.NPR))$$

APA004.TES = 1 if the provider is classified as a 'Total exemption SMALL' body by Companies House, 0 otherwise

PRV001.POS = 1 if the outcome of the provider's previous review is positive, 0 otherwise

PRV004.NPC =1 if the provider has had no previous comparable review, 0 otherwise

PRV006.NPR = 1 if the provider has no previous 'worst judgement' as they have no previous comparable review, 0 otherwise

PRV009.NPR = 1 if the provider has never previously received a judgement concerning teaching and learning, 0 otherwise

PRV010.NPR = 1 if the provider has not received a judgment concerning teaching and learning in a previous comparable review, 0 otherwise

PRV012.NPR = 1 if the provider has never previous been reviewed in relation to the provision of information, 0 otherwise

$$P(Requires\ improvement\ to\ meet\ UK\ expectations) = \frac{e^B}{1+e^B}$$

Where, to one significant figure:

$$B = (-0.7 + (0.7 \times APA004.SMA) - (0.002 \times PRV001.POS) - (0.08 \times APA002) + (0.0003 \times APA014\_Ca1i) + (2.9 \times CON003))$$

APA004.SMA = 1 if the provider is classified as a 'SMALL' body by Companies House, 0 otherwise

$$P(Meets\ UK\ expectations) = \frac{e^C}{1+e^C}$$

Where, to one significant figure:

$$C = (2.4 + (0.0002 \times APA001) + (3.5 \times APA004.FUL) - (1.7 \times APA004.TES) + (0.0004 \times APA009) + (0.001 \times APA011) - (0.0001 \times APA016) + (0.6 \times CON002) + (1.8 \times PRV002.YES) + (1.E-14 \times PRV003.YES)$$

APA004.FUL = 1 if the provider is classified as a 'FULL' body by Companies House, 0 otherwise

PRV003.YES = 1 if the provider has ever received a negative review, 0 otherwise

The large number of metrics contained in the three models makes their interpretation far from intuitive. The main effects are that being classified as a 'Total exemption SMALL' company for

reporting purposes by Companies House and having never had a previous, comparable review greatly increase the likelihood of being forecast 'Does not meet UK expectations' whereas having had a positive previous comparable review reduces the likelihood. The likelihood of being forecast 'Requires improvement to meet UK expectations' increases when classified as 'SMALL' by Companies House, *QAA concerns* have been upheld in relation to *academic standards* at the provider, and, somewhat counterintuitively, when a provider has seen an increase in their net assets since their last set of accounts was submitted. Conversely, the likelihood of being judged 'Requires improvement to meet UK expectations' is reduced by having a positive previous review outcome and, albeit by a very small amount, having outstanding mortgage charges. Finally, the chances of a provider being forecast to be 'Meets UK expectations' are increased by being large with a strong financial position, having had invalid *QAA concerns* raised against them, and having had a negative outcome for their last review. Perhaps more logically than for the other two models, the chances of being forecast to be 'Meets UK expectations' are reduced by being small and having large payments to make within one year.

7.3.3. Evaluating the Model

Evaluating the model becomes far more challenging when we are attempting to predict the likelihood of three outcomes, i.e. 'Meets UK expectations', 'Requires improvement to meet UK expectations', or 'Does not meet UK expectations', rather than two, i.e 'satisfactory' or 'unsatisfactory'. To understand why, it is easiest to consider the practical challenges of implementing a risk-based approach centred on a model predicting the likelihood of three outcomes occuring. The first challenge is defining what is a successful prediction? Should one class predicting 'Does not meet UK expectations' as a poor outcome when the result is that the provider actually 'Requires improvement to meet UK expectations'? Similarly, how should the providers be prioritised? If one provider has a 30% predicted likelihood of being 'Requires improvement to meet UK expectations' and a 30% predicted likelihood of being 'Does not meet UK expectations' should that be prioritised over an second provider with a 40% predicted likelihood of being 'Does not meet UK expectations' and 60% likelihood of being 'Meets UK expectations'? The former has the higher overall likelihood of being 'unsatisfactory' whereas the latter has a higher likelihood of being the most severe level of 'unsatisfactory'. Such practical, normative challenges translate to the model evaluation. An ROC curve is based upon correct and incorrect predictions, if a correct prediction cannot be defined then an ROC curve cannot be developed. If one defines a correct prediction as the the most likely outcome occurs, then a model which successfully predicts a large number of providers will be 'unsatisfactory' but gets the 'Does not meet UK expectations' / 'Requires improvement to meet UK expectations' classification wrong will be evaluated as worse

than a model which does a far worse job of predicting 'satisfactory' versus 'unsatisfactory' providers, but correctly identifies the precise class when it does get them correct.

7.3.3.1. Testing the Fit of the Model

As Figure 7.15 below illustrates, visualising the model predictions and review outcomes for our representative training data set of 211 reviews is far more challenging when considering three dimensions compared to two:

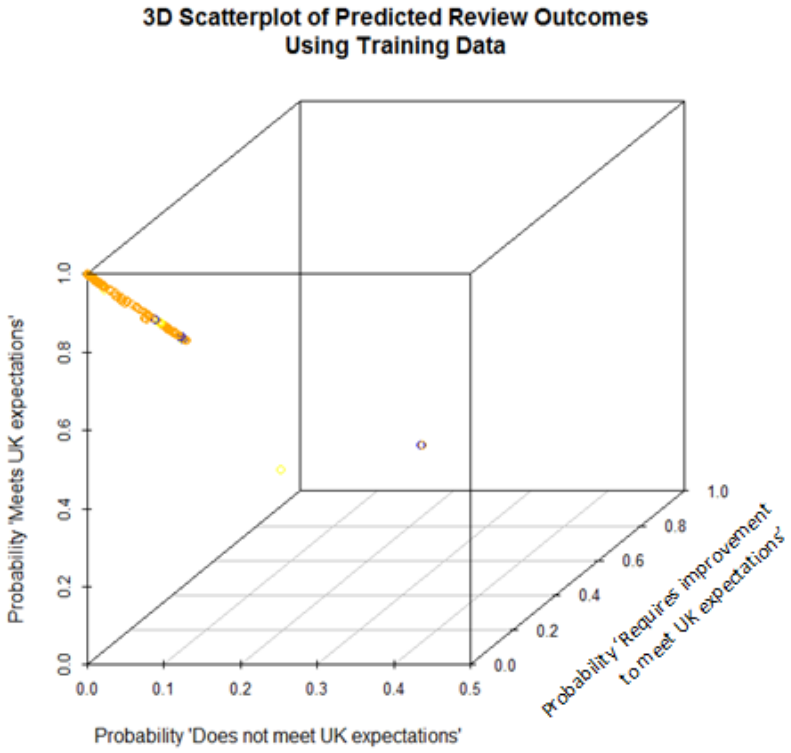

Figure 7.15: A 3D plot of each provider's review outcome and predicted probability of being judged 'Does not mee UK expectations', 'Requires improvement to meet UK expectations' or 'Meets UK expectations'. Each circle indicates and individual alternative provider and the blue, yellow or orange colour of the circle indicate the provider was judged 'Does not meet UK expectations', 'Requires improvement to meet UK expectations', or 'Meets UK expectations' respectively.

All reviews except two are clustered in the top-left corner of the plot with a predicted probability that they will be judged 'Meets UK expectations' of greater than 80%. The narrow range of predicted proababilities in the best possible model suggests a weak relationship between the metrics and the categorical outcome of the *academic standards* question of QAA reviews.

Clearly interpreting the model with three possible outcomes, and hence in three dimensions, is challenging (and would be even more so if the fourth 'Commended' outcome was applicable to the *academic standards* question). A more straightforward way to evaluate how well the model fits the data with which it was developed is to see how many reviews would have been required

to succesfully prioritise the reviews of all providers who are judged either ‘Does not meet UK expectations’ or ‘Requires improvement to meet UK expectations’. The prioritisation can realistically be done in two ways. The reviews could be prioritised either by the combined probability of being judged ‘Does not meet UK expectations’ or ‘Requires improvement to meet UK expectations’, or simply by the probability of being judged ‘Does not meet UK expectations’. Both of these options are explored in turn below.

Figure 7.18 and Table 7.5 below show that when the reviews are priotrised by the combined probability of a provider being judged either ‘Does not meet UK expectations’ or ‘Requires improvement to meet UK expectations’, 101 out of the 211 reviews in the training data set would have been required to identify all the providers judged either ‘Does not meet UK expectations’ or ‘Requires improvement to meet UK expectations’. Whether one views success in terms of identifying all ‘unsatisfactory’ provision or the less stringent identification of a high-risk group containing a greater proportion of the ‘unsatisfactory’ providers than a low-risk group, the model performs reasonably well. The range of predicted probabilities is again very narrow and the lack of certainty in the model is reflected in another high error rate: 79 out of the 101 (78%) providers prioritised would have been reviewed unnecessarily but this would still have meant – if operating with perfect hindsight and the ability to stop reviewing immediately after the last ‘unsatisfactory’ provider had been identified - the QAA could have avoided 110 unnecesary reviews. Everyone of the 22 ‘unsatisfactory’ providers were in the riskiest 50% of providers.

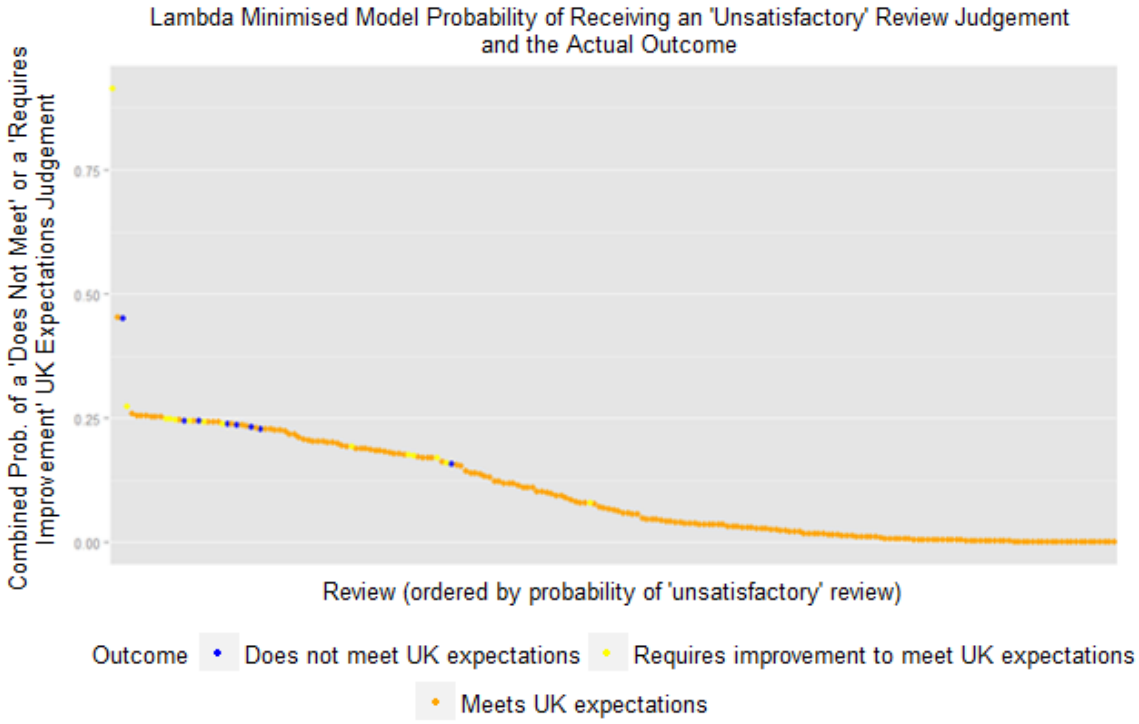

Figure 7.16: Predicted probabilities for each of the 211 complete, comparable reviews used to train the model, ordered by the combined proability of being judged either 'Does not meet UK expectations' or 'Requires improvement to meet UK expectation', and their actual outcome.

Predicted Probability of Outcome				Number of Judgements		
Does not meet' + 'Requires improvement'	Does not meet UK expectations	Requires improvement to meet UK expectations	Meets UK expectations	Does not meet UK expectations	Requires improvement to meet UK expectations	Meets UK expectations
91.28%	0.08%	91.21%	8.72%	0	1	0
45.14%	44.37%	0.76%	54.86%	1	1	1
27.30%	8.30%	18.99%	72.70%	1	2	1
24.92%	9.40%	15.52%	75.08%	1	3	8
24.90%	9.63%	15.27%	75.10%	1	4	8
24.59%	9.51%	15.08%	75.41%	1	5	8
24.52%	9.51%	15.01%	75.48%	2	5	9
24.47%	9.43%	15.04%	75.53%	2	6	9
24.36%	9.46%	14.89%	75.64%	3	6	10
24.28%	9.41%	14.87%	75.72%	3	7	10
23.88%	9.29%	14.59%	76.12%	3	8	13
23.79%	8.68%	15.12%	76.21%	4	8	13
23.62%	9.33%	14.29%	76.38%	5	8	14
23.12%	8.94%	14.18%	76.88%	6	8	16
22.86%	8.82%	14.04%	77.14%	7	8	17
19.18%	3.82%	15.36%	80.82%	7	9	35
17.54%	6.77%	10.76%	82.46%	7	10	46
17.39%	6.73%	10.66%	82.61%	7	11	46
16.95%	6.57%	10.38%	83.05%	7	12	50
15.86%	5.76%	10.10%	84.14%	7	13	51
15.65%	6.04%	9.61%	84.35%	8	13	51
7.96%	0.59%	7.37%	92.04%	8	14	79

Table 7.5: The number of 'Does not meet UK expectations', 'Requires improvement to meet UK expectations' and 'Meets UK expectations'judgement that would have resulted from decreasing the threshold – defined as the probability of being judged either 'Does not meet UK expectations' or 'Requires improvement to meet UK expectations' – required to prompt an alternative provider review.

Figure 7.17 and Table 7.6 below show that prioritising reviews based on the predicted probability of a provider being judged 'Does not meet UK expectations' – a feasible approach for any oversight body whose priority is to visit the worst providers first, then visit the less severe 'Requires improvement to meet UK expectations' providers next – yields slightly worse results that prioritising the providers based on the combined probability. Due to two outliers, 183 reviews would need to be prioritised in order to review all of the 'unsatisfactory' providers. In this scenario 161 out of the 183 (88%) providers prioiritised would have been reviewed unnecessarily. Under this prioritisation approach the eight providers judged 'Does not meet UK expectations' would have all been reviewed after 63 reviews, marginally fewer than the 72 reviews under the first

approach. 20 of the 22 unsatisfactory providers would have still been prioritised as part of the 50% of riskiest providers however.

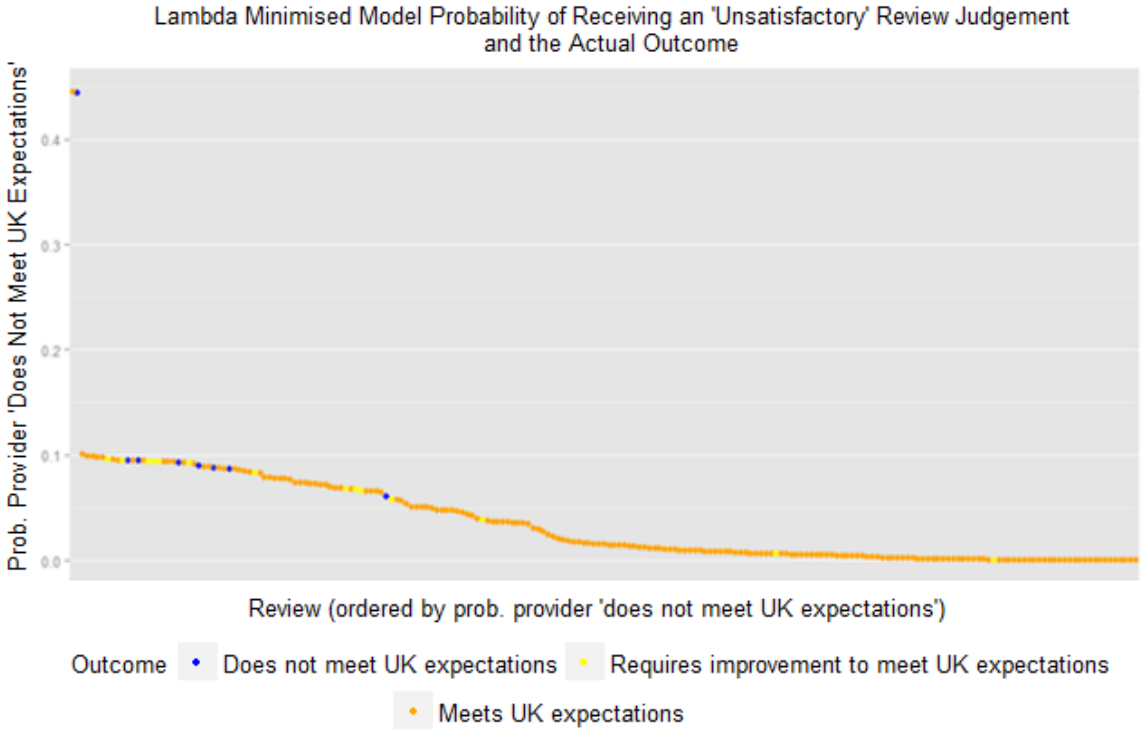

Figure 7.17: Predicted probabilities for each of the 211 complete, comparable reviews used to train the model, ordered by the combined proability of being judged 'Does not meet UK expectations', and their actual outcome.

Predicted Probability of Outcome			Number of Judgements		
Does not meet UK expectations	Requires improvement to meet UK expectations	Meets UK expectations	Does not meet UK expectations	Requires improvement to meet UK expectations	Meets UK expectations
44.37%	0.76%	54.86%	1	0	1
9.63%	15.27%	75.10%	1	1	6
9.51%	15.08%	75.41%	1	2	8
9.51%	15.01%	75.48%	2	2	8
9.46%	14.89%	75.64%	3	2	9
9.43%	15.04%	75.53%	3	3	10
9.41%	14.87%	75.72%	3	4	10
9.40%	15.52%	75.08%	3	5	10
9.33%	14.29%	76.38%	4	5	13
9.29%	14.59%	76.12%	4	6	14
8.94%	14.18%	76.88%	5	6	15
8.82%	14.04%	77.14%	6	6	17
8.68%	15.12%	76.21%	7	6	19
8.30%	18.99%	72.70%	7	7	23
6.77%	10.76%	82.46%	7	8	40

6.73%	10.66%	82.61%	7	9	41
6.57%	10.38%	83.05%	7	10	41
6.04%	9.61%	84.35%	8	10	45
5.76%	10.10%	84.14%	8	11	45
3.82%	15.36%	80.82%	8	12	62
0.59%	7.37%	92.04%	8	13	119
0.08%	91.21%	8.72%	8	14	161

Table 7.6: The number of 'Does not meet UK expectations', 'Requires improvement to meet UK expectations' and 'Meets UK expectations'judgment that would have resulted from decreasing the threshold – defined as the probability of being judged 'Does not meet UK expectations' – required to prompt an alternative provider review.

Interestingly, regardless of how the more granular predictions are used to prioritise reviews, the model cannot effectively differentiate the providers judged 'Does not meet UK expectations' and the providers judged 'Requires improvement to meet UK expectations'. This defeats the main purporse of the more granular probabilities: prioritising of the more serious 'Does not meet UK expectations' cases first. No matter which probability, or combination of probabilities, is used to prioritise reviews it is again the case that the distribution of predicted probabilities is very flat with only one provider having a predicted probability of being 'unsatisfactory' greater than 50%. Despite the lack of certainty in the model's predictions, and the high error rate associated with it, all the 'unsatisfactory' providers would have been prioritised in the first 50% of reviews when the combined probability was used to determine the order of reviews. Whether this was good fortune, or the result of genuine underlying relationships between the metrics and review outcomes, will be made clearer by applying the model to the test data.

7.3.3.2. Assessing the Model's Predictions

To test the *academic standards* model 30% of the reviews were held back. Figure 7.18 and Table 7.7 show how the model performed on the test data when the reviews are ordered by the combined probability of a provider being judged either 'Does not meet UK expectations' or 'Requires improvement to meet UK expectations':

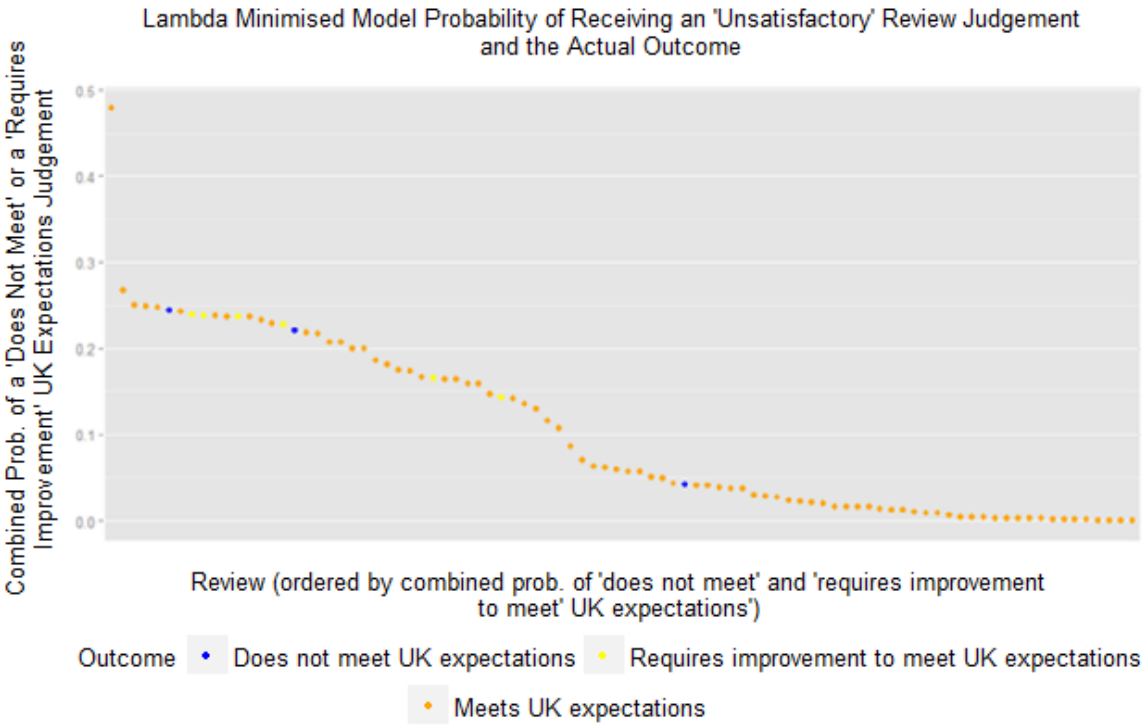

Figure 7.18: Predicted probabilities for each of the 90 complete, comparable reviews used to test the model, ordered by the combined proability of being judged 'Does not meet UK expectations', and their actual outcome.

Predicted Probability of Outcome				Number of Judgements		
Does not meet' + 'Requires improvement'	Does not meet UK expectations	Requires improvement to meet UK expectations	Meets UK expectations	Does not meet UK expectations	Requires improvement to meet UK expectations	Meets UK expectations
24.43%	9.47%	14.96%	75.57%	1	0	5
23.95%	9.60%	14.35%	76.05%	1	1	6
23.77%	9.20%	14.57%	76.23%	1	2	6
23.60%	9.30%	14.30%	76.40%	1	3	8
22.75%	8.80%	13.95%	77.25%	1	4	11
22.00%	8.52%	13.48%	78.00%	2	4	11
16.55%	6.10%	10.45%	83.45%	2	5	22
14.27%	2.95%	11.32%	85.73%	2	6	27
4.13%	1.53%	2.60%	95.87%	3	6	42

Table 7.7: The number of 'Does not meet UK expectations', 'Requires improvement to meet UK expectations' and 'Meets UK expectations'judgement that would have resulted from decreasing the threshold – defined as the probability of being judged 'Does not meet UK expectations' – required to prompt an alternative provider review.

The same pattern that was witnessed when assessing the fit of the model is apparent in the test data. This matching pattern provides reassurance that the model has identified true patterns in the data without also incorporating the statistical noise and random variation. When prioritised in order of the combined probability of being judged either 'Does not meet UK expectations' or

'Requires improvement to meet UK expectations' all 'unsatisfactory' providers would have been reviewed by the point 51 of the 90 reviews had taken place. This means that, had the QAA conducted these reviews and stopped at the optimal point 41 reviews of 'satisfactory' providers could have been avoided. Of the 51 reviews that would have taken place however, 42 (82%) would have been of 'satisfactory' providers. All but one of the 'unsatisfactory' providers were in the riskiest half of providers. Whilst the model struggles to accurately predict whether individual providers will be 'unsatisfactory', it is able to identify a high-risk group which will contain the 'unsatisfactory' providers along with four times as many 'satisfactory' providers. Seemingly, all 'unsatisfactory' providers share certain characteristics, but only 20% of providers with those characteristics are 'unsatisfactory'. As with each model in this chapter two further tests – assessing the model's predictions with the new data and at a specific point in time - would be beneficial but were not possible.

7.3.3.3. Binary Model

The fact that the granular model's best predictions come from combining the probabilities of being 'Does not meet UK expectations' and 'Requires improvement to meet UK expectations', and that model fails to effectively differentiate between providers judged 'Does not meet UK expectations' and providers judged 'Requires improvement to meet UK expectations', raises an obvious question: is the more complex to interpret and maintain granular model worth the additional effort over the binary 'satisfactory' / 'unsatisfactory' model?

The optimal binary model for the *academic standards* question contained six metrics:

- APA001 - Age at time of review.
- APA004 – Financial Accounts Type (e.g. 'Total exemption', 'Full').
- APA011 – Amount of Cash At Bank And In Hand (£000's).
- PRV002 - Has the provider been reviewed and received a negative outcome in the last review?
- CON002 - Count of *QAA concerns* raised which do not relate to quality, standards, information or enhancement (and are therefore automatically invalid).
- CON003 - Count of *QAA concerns* upheld since previous review which relate to academic standards

As expected there is some overlap with the model for the overall review outcome. The binary *academic standards* model contains three of the four metrics that comprise the overall model. The specific probability of each provider being judged 'unsatisfactory' is given by:

$$P(Unsatisfactory) = \frac{e^A}{1 + e^A}$$

Where:

$$A = \big(-2.0 - (0.0002 \times APA001) + (0.92 \times APA004.TES) - (0.99 \times APA004.FUL)\big) - (0.0005 \times APA011) - (1.72 \times PRV001.POS) - (0.62 \times CON002) + (2.68 \times CON003))$$

APA004.TES = 1 if the provider is classified as a 'Total exemption SMALL' body by Companies House, 0 otherwise

APA004.FUL = 1 if the provider is classified as a 'FULL' body by Companies House, 0 otherwise

PRV001.POS = 1 if the outcome of the provider's previous review is positive, 0 otherwise

The model has an impressive ROC 'area under the curve' value of 0.842. Figure 7.19 and Table 7.8 below show the performance of the model on the data with which it was developed:

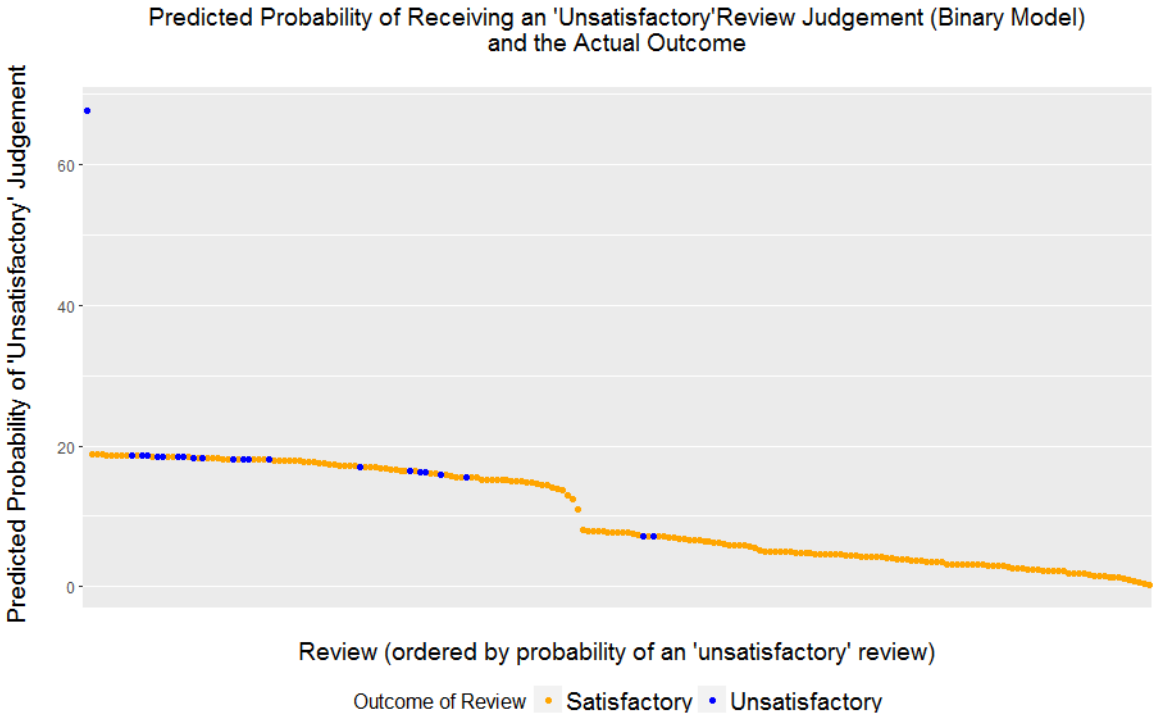

Figure 7.19: Predicted probabilities for each of the 211 complete, comparable reviews used to train the model and their actual outcome.

	Predicted Probability of 'Unsatisfactory' Review	Number of 'Unsatisfactory' Reviews	Number of 'Satisfactory' Reviews	Accuracy rate
← Decreasing probability of a review being 'unsatisfactory' required to trigger inspection	67.62%	1	0	0.90
	18.52%	5	10	0.87
	18.35%	9	13	0.88
	18.10%	13	20	0.86
	16.33%	17	50	0.74
	15.53%	20	56	0.73
	7.25%	21	90	0.57
	7.15%	22	91	0.57

Table 7.8: The number of 'satisfactory' and 'unsatisfactory' alterntative provider reviews that would have resulted from decreasing the threshold required to prompt a review (only select points are shown).

Considering the outcome of the *academic standards* question at a binary level results in very similar performance. When prioritised in order of the predicted likelihood of being 'unsatisfactory', 113 out of the 211 reviews in the training data set would have been required to identify all the 'unsatisfactory' providers. This is only 12 more than for the more granular model and identifying all 'unsatisfactory' provision despite being far simpler. 91 out of the 113 (80.5%) providers prioritised would have been reviewed unnecessarily which is a marginally higher error rate than for the more granular model (78%). All but two of the 22 'unsatisfactory' providers were in the riskiest 50% of providers.

As shown in Figure 7.20 and Table 7.9 below, the model performs similarly when applied to the withheld testing data: all but one of the 'unsatisfactory' reviews are in the top-third of riskiest providers.

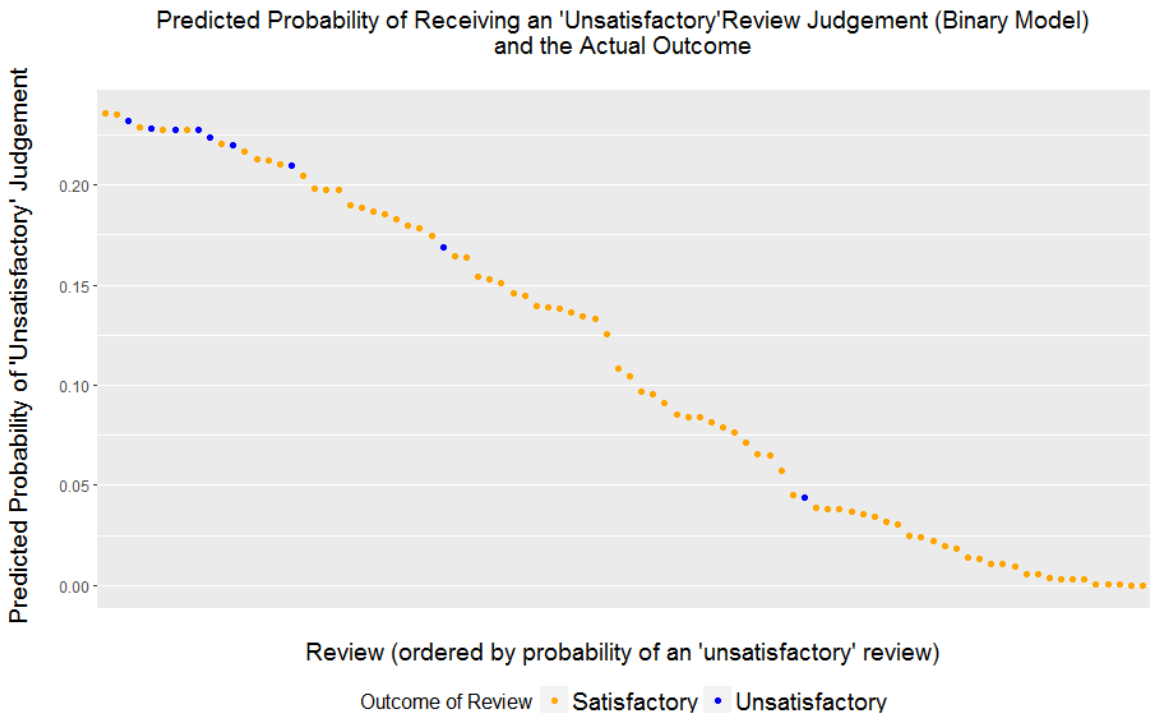

Figure 7.20: Predicted probabilities for each of the 90 complete, comparable reviews in the withheld test data set and their actual outcome.

	Predicted Probability of 'Unsatisfactory' Review	Number of 'Unsatisfactory' Reviews	Number of 'Satisfactory' Reviews	Accuracy rate
← Decreasing probability of a review being 'unsatisfactory' required to trigger inspection	23.20%	1	2	0.89
	22.84%	2	3	0.89
	22.79%	3	4	0.89
	22.73%	4	5	0.89
	22.36%	5	5	0.90
	21.97%	6	6	0.90
	20.95%	7	10	0.87
	16.91%	8	22	0.74
	4.40%	9	52	0.42

Table 7.9: The number of 'satisfactory' and 'unsatisfactory' alterntative provider reviews that would have resulted from decreasing the threshold required to prompt a review (only select points are shown) when applied to the testing data.

The more granular model is therefore only marginally more accurate than the binary 'satisfactory' / 'unsatisfactory' model for the *academic standards* review question. Despite the more granular model being far more complex to interpret and requiring more effort to maintain, it fails to succesfully differentiate between providers judged to be 'Does not meet UK expectation' or 'Requires improvement to meet UK expectations' to allow for the better targetting of resource: the core aim of the more granular model.

7.3.3.4. Summary

The individual metrics that were significant predictors of *academic standards* followed a now familiar pattern and demonstrated which providers were most likely to be 'satisfactory' but were of limited use in identifying providers judged either 'Does not meet UK expectations' or 'Requires improvement to meet UK expectations'. Both the λ_{min} and λ_{1se} granular models were complicated to interpret and intially appeared to slightly overfit the data with the inclusion of metrics with questionable coefficients. The λ_{min} model explored in this section contained 16 metrics covering each of the three key data sources: finance, *QAA concerns* and previous review performance.

When the reviews were prioritised based on the combined probability of being judged 'Does not meet UK expectations' or 'Requires improvement to meet UK expectations' all such providers were prioritised in the riskiest half of providers. Promisingly the same was true when the model was applied to the test data not used in the development of the model. The model had a high error rate which suggests that all 'unsatisfactory' providers share some charactertistics, but those characteristics are shared by four times as many 'satisfactory' providers making accurate prioritisation impossible. Furthermore, the model struggled to successfully differentiate between those providers that were judged 'Does not meet UK expectations' and those that were judged 'Requires improvement to meet UK expectations'.

The answer to the question posed at the start of this section:

> *Using naturally-complete metrics, could the exact outcome of QAA reviews of academic standards at alternative providers have been successfully predicted?*

is no: the exact 'Does not meet UK expectations', 'Requires improvement to meet UK expectations' and 'Meets UK expectations' outcomes could not be predicted. On a positive note, the binary 'satisfactory'/ 'unsatisfactory' outcome for the *academic standards* review question was better predicted by both the granular and binary models.

The fact that the question-level analysis was marginally more accurate than the review-level model suggests that it is worth exploring the data at question level for the two remaining areas where there is sufficient data: *teaching and learning* and *the provision of information.* It has already been discussed that there are significant normative challenges working with predicted probabilities of outcomes as granular as question (e.g. *academic standards, teaching and learning,* etc.) and specific outcome-level (e.g. 'Does not meet UK expectations', 'Requires improvement to meet UK expectations'). The fact that best results come from combining the predicted probabilities of the provider being judged 'Does not meet UK expectations' or 'Requires

improvement to meet UK expectations', coupled with the challenges of operating an approach at a finer level, suggests that the best way forward is to assess the two remaining questions but at the binary 'satisfactory'/ 'unsatisfactory' level.

7.4. Results – Teaching and Learning

Using naturally-complete metrics, could the aggregated outcome of QAA reviews of teaching and learning at alternative providers have been successfully predicted?

7.4.1. Initial Data Exploration

The first step in the analysis was to examine how similar the outcome of the *teaching and learning* question was to the overall outcome of reviews of alternative providers.

	Unsatisfactory		Satisfactory	
	Does not meet UK expectations	Requires improvement to meet UK expectations	Meets UK expectations	Commended
Overall review outcome	41		264	
Teaching and Learning outcome	10	18	276	1

Table 7.10: The results for the overall outcome and teaching and learning section of each alternative provider review.

With 28 out of 41 providers judged 'unsatisfactory' overall having also been judged 'unsatisfactory' in relation to *teaching and learning* it is likely the models for predicting the outcome of the overall review and the outcome of the *teaching and learning* question will be similar.

The second step in the analysis was to explore which individual metrics had a strong relationship with the 'satisfactory' or 'unsatisfactory' outcome of the *teaching and learning* review question. The only metrics with a p-value less than 0.25 were financial metrics developed from the providers' accounts:

Metric Code	Metric Description	P-value
APA001	Age at time of review	0.035
APA002	Number of outstanding mortgage charges	0.140

Table 7.11: A breakdown of all metrics from the teaching and learning data set with a p-value of less than 0.25.

Two significant metrics is fewer than there were for the outcome of the review overall and for the *academic standards* question. The provider's age at the time of the review was a significant metric in both previous analyses and makes intuitive sense when thought of as a proxy for experience. The *APA002 – Number of outstanding mortgage charges* metric was not a significant individual

predictor of either the overall review outcome or the *academic standards* question. Figure 7.21 below shows that the metrics' significance in relation to *teaching and learning* is a quirk of the data: providers with a greater number of outstanding mortgage charges have in the past been less likely to be 'unsatisfactory'. This makes little intuitive sense. There is no obvious reason why those providers with fewer existing debts would be more likely to be 'unsatisfactory' in relation to *teaching and learning*. Having four or more outstanding mortgage charges is rare, and of those few providers that have four or more, none have been found unsatisfactory' with regards *teaching and learning.*

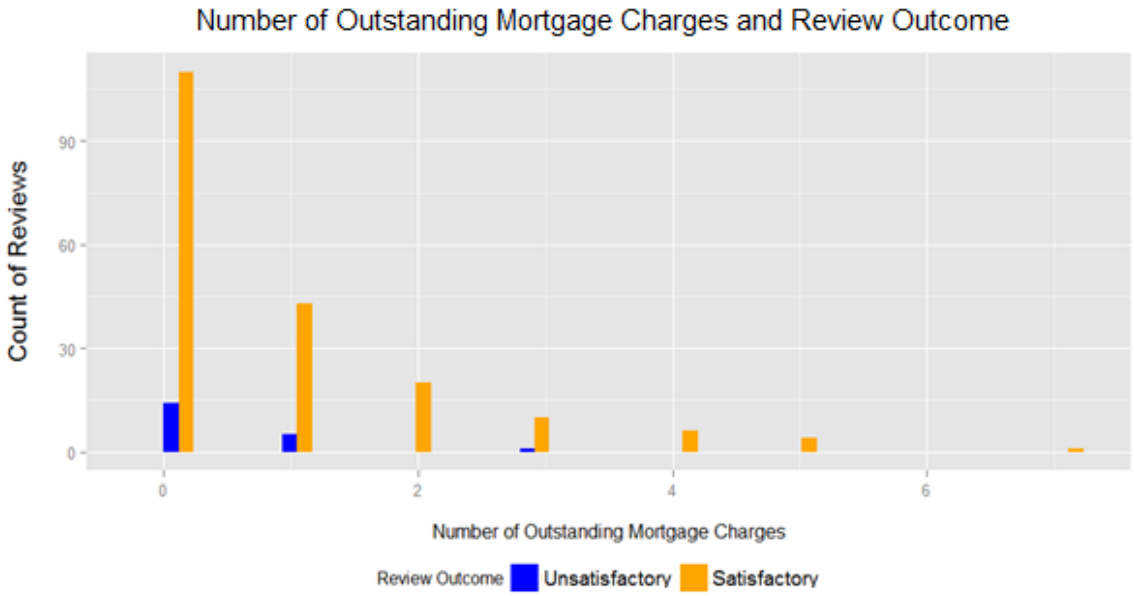

Figure 7.21: Each provider's count of outstanding mortgage charges at the time of their review and the outcome of the teaching and learning question of that review.

7.4.2. Fitting the Model

Running the *elastic net* procedure we obtain the diagnostic plots shown below.

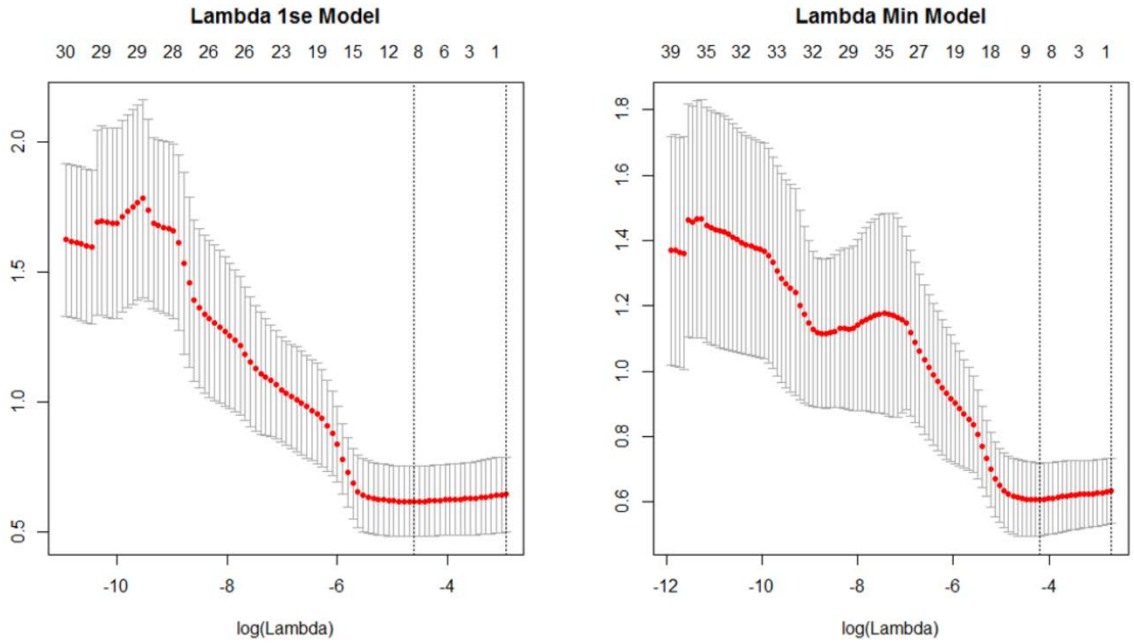

Figure 7.22: The diagnostic plots for the λ_{1se} (left) and λ_{min} (right) model for the binary teaching and learning model.

The λ_{min} and λ_{1se} models are very similar and neither seem to overfit the data so the preference is for the more accurate λ_{min} model. This model contains six metrics including both significant metrics discussed above and four of the six metrics that make up the binary *academic standards* model.

- APA001 - Age at time of review.
- APA002 – Count of the number of outstanding mortgage charges against the provider.
- APA004 – Financial Accounts Type (e.g. 'Total exemption', 'Full').
- PRV002 - Has the provider been reviewed and received a negative outcome in the last review?
- PRV003 - Has ever received a negative review?
- CON002 - Count of *QAA concerns* raised which do not relate to quality, standards, information or enhancement (and are therefore automatically invalid).

The specific λ_{min} model calculates the probability of an unsatisfactory review as:

$$P(Unsatisfactory) = \frac{e^A}{1 + e^A}$$

where:

A = -2.08 + (-1.28 x APA004.FUL) + (-0.11 x APA004.GRP) + (0.80xAPA004.SMA) + (0.47 x APA004.TES) + (-0.81 x PRV002.YES) + (-0.43 x PRV003.YES) + (-0.00006 x APA001) + (-0.14 x APA002) + (-0.63 x CON002)

and:

APA004.FUL = 1 if the provider is classified as a 'FULL' body by Companies House, 0 otherwise
APA004.GRP = 1 if the provider is classified as a 'GROUP' body by Companies House, 0 otherwise
APA004.SMA = 1 if the provider is classified as a 'SMALL' body by Companies House, 0 otherwise
APA004.TES = 1 if the provider is classified as a 'Total exemption SMALL' body by Companies House, 0 otherwise
PRV002.YES = 1 if the provider been reviewed and received a negative outcome in the last review, 0 otherwise
PRV003.YES =1 if the provider has ever received a negative review, 0 otherwise

As the coefficients for the APA004.SMA and APA004.TES metrics are positive, the likelihood of receiving an 'unsatisfactory' review will increase for small providers. Conversely, the likelihood will decrease for providers classed as 'Full' or 'Group' and also, counterintuitively, for those providers that have either previously been reviewed and received a negative outcome in the last review, have ever received a negative review, or have ever been the subject of an invalid QAA Concern. Each of these metrics and their possible links with quality assurance processes have been discsussed in previous sections and so will not be discussed again here.

7.4.3. Evaluating the Model

7.4.3.1. Testing the Fit of the Model

Figure 7.24 below shows the ROC curve for this model when applied to the data used to develop it. The 'area under the curve' value of 0.753 suggests a reasonable rate of 'unsatisfactory' alternative providers being succesfully prioritised as the threshold criteria for triggering a review is lowered. The shape of the 'curve', with the majority of the area under the curve being created on the righthand side of the plot, indicates a high proportion of early predictions were incorrect prior to a large number of true negatives (correct predictions of 'satisfactory provision) beyond a certain point.

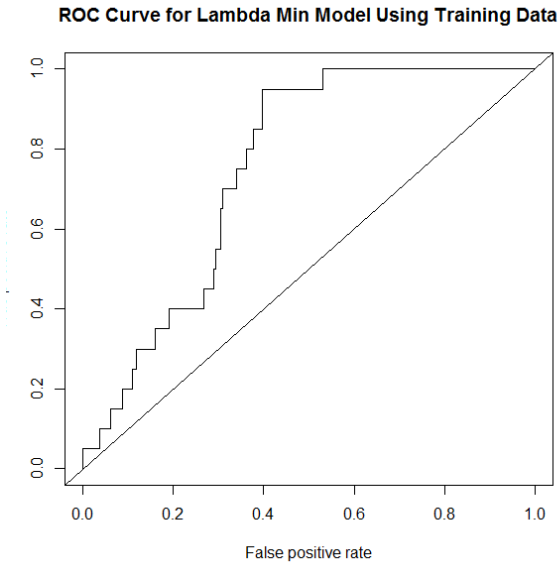

Figure 7.23: The ROC curve for the λ_{min} model fitted to the training data set for teaching and learning.

Figure 7.24 and Table 7.10 below show the effect of lowering the threshold required for the model's predicted probability of a provider being 'unsatisfactory' to trigger a review and confirm the pattern suggested above by the ROC curve.

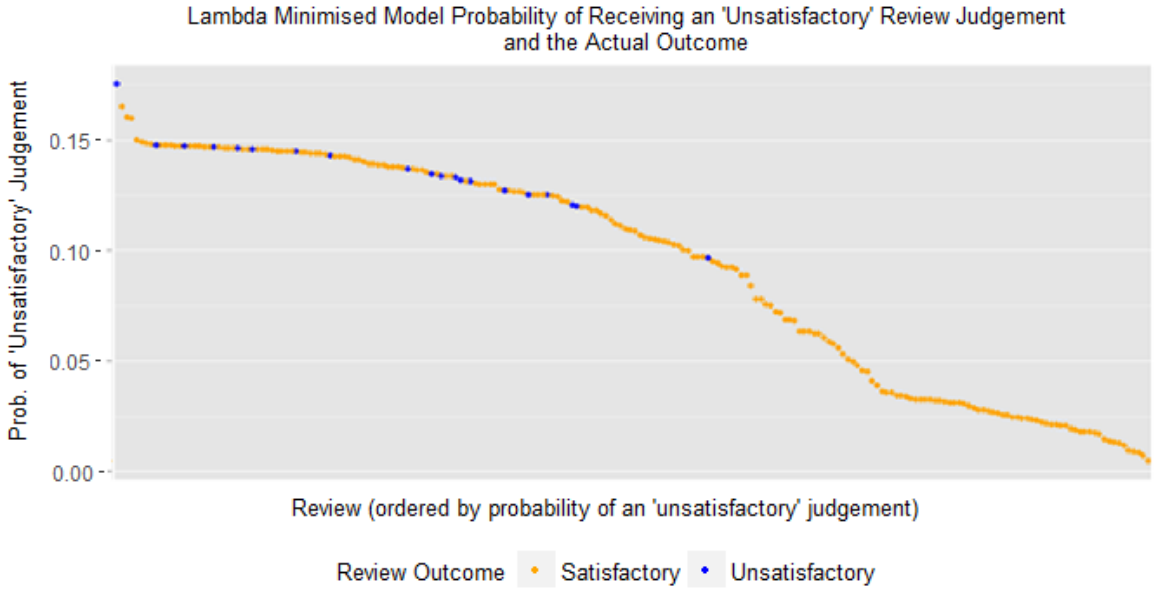

Figure 7.24: Predicted probabilities for each of the 215 complete, comparable teaching and learning review questions used to train the model and their actual outcome.

	Predicted Probability of 'Unsatisfactory' Review	Number of 'Unsatisfactory' Reviews	Number of 'Satisfactory' Reviews	Accuracy rate
← Decreasing probability of a review being 'unsatisfactory' required to trigger inspection	17.50%	1	0	0.91
	14.66%	4	17	0.85
	14.30%	8	37	0.77
	13.30%	12	59	0.69
	12.51%	16	70	0.66
	9.64%	20	103	0.52

Table 7.12: The number of 'satisfactory' and 'unsatisfactory' reviews that would have resulted from decreasing the threshold required to prompt a review concerning teaching and learning.

The model performs relatively poorly to begin with prioritsing a number of 'satisfactory' providers amongst the 'unsatisfactory' providers but a collection of 'unsatisfactory' providers in the second quartile means that all but one 'unsatisfactory' provider are prioritised in the top 45% of riskiest providers. Once the one outlying provider has been prioritised all 'unsatisfactory' providers would be categorised in the top 57.5% of providers by risk. Had the system been implemented perfectly with the review activity being ceased immediately after the final 'unsatisfactory' review had been undertaken this would have resulted in an error rate of 103 / (20 + 103) = 83.7%. Whilst this figure is unlikely to sit well with alternative providers categorised as high risk, it could have prevented

92 reviews of 'satisfactory' providers. Indeed, all bar one of the 'unsatisfactory' reviews resided in the riskiest 50% of providers. Once more the narrow range of predicted probabilities, the subsequent high error rate, and the fact that that all bar one 'unsatisfactory' review was in the riskiest 50% of providers suggests there are characteristics shared by 'unsatisfactory' providers, but unfortunately these characteristics are shared by a far greater number of 'satisfactory' providers.

7.4.3.2. Assessing the Model's Predictions

More important than how well the model fits the data with which it was developed is how well it performs when making new predictions. The model performs poorly. The ROC curve which passes below the 45 degree line towards the end indicating that – at that point – the QAA would be better off doing the opposite of what the model suggests. The curve does recover and finish above the 45 degree line; however, the low area under the curve value of 0.610 indicates that the model is ineffective.

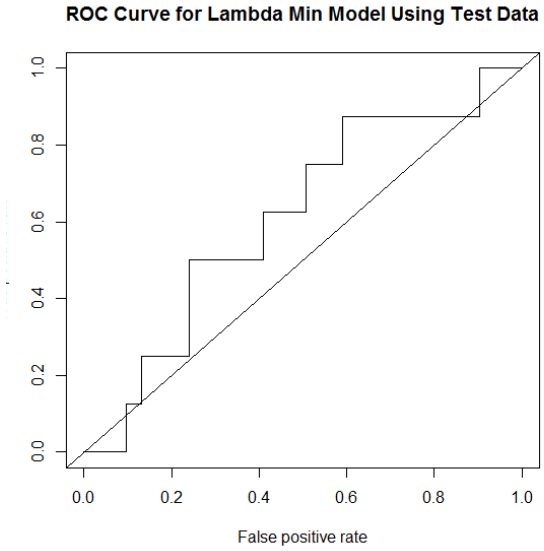

Figure 7.25: The ROC curve for the λ_{min} model fitted to the test data set.

Figure 7.26 and Table 7.11 below show the model's application to the training set comprising 90 reviews:

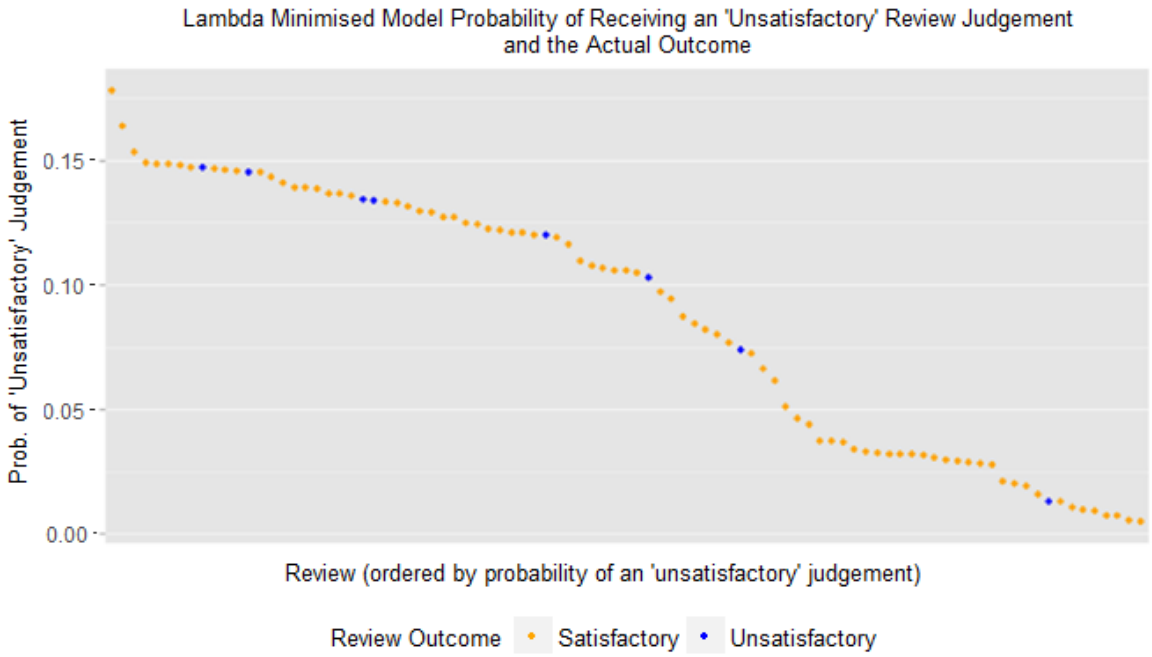

Figure 7.26: The predicted probabilities and actual outcomes of the teaching and learning question for reviews contained in the testing set.

Predicted Probability of 'Unsatisfactory' Review	Number of 'Unsatisfactory' Reviews	Number of 'Satisfactory' Reviews
14.69%	1	8
14.53%	2	11
13.40%	3	20
13.37%	4	20
11.97%	5	34
10.29%	6	42
7.38%	7	49
1.31%	8	75

Table 7.13: The number of 'satisfactory' and 'unsatisfactory' reviews that would have resulted from decreasing the threshold required to prompt a review concerning teaching and learning.

Although half of the 'unsatisfactory' providers would have been identified having undertaken just over a quarter – 24 out of 90 – reviews, to have identified the remaining half of 'unsatisfactory' providers nearly all - 83 of the 90 - providers would have had to have been reviewed. Just 5 out of 8 of the 'unsatisfactory' providers were amongst the riskiest 50% of providers.

The relationships between the selected metrics and the outcome of the *teaching and learning* review question suggested by the model were seemingly just chance relations in the subset of data used to develop the model rather than true relations.

7.4.3.3. Summary

The best model calculated the predicted probability of an alternative provider being judged 'unsatisfactory' in relation to their *teaching and learning* based on previous review performance, the age, size and financial position of the provider, and the number of invalid *QAA concerns* raised against them. The model determined the predicted likelihood of being judged 'unsatisfactory' as:

$$P(Unsatisfactory) = \frac{e^A}{1 + e^A}$$

where:

A = -2.08 + (-1.28 x APA004.FUL) + (-0.11 x APA004.GRP) + (0.80xAPA004.SMA) + (0.47 x APA004.TES) + (-0.81 x PRV002.YES) + (-0.43 x PRV003.YES) + (-0.00006 x APA001) + (-0.14 x APA002) + (-0.63 x CON002)

and:

APA004.FUL = 1 if the provider is classified as a 'FULL' body by Companies House, 0 otherwise
APA004.GRP = 1 if the provider is classified as a 'GROUP' body by Companies House, 0 otherwise
APA004.SMA = 1 if the provider is classified as a 'SMALL' body by Companies House, 0 otherwise
APA004.TES = 1 if the provider is classified as a 'Total exemption SMALL' body by Companies House, 0 otherwise
PRV002.YES = 1 if the provider been reviewed and received a negative outcome in the last review, 0 otherwise
PRV003.YES =1 if the provider has ever received a negative review, 0 otherwise

Whilst the model was a good fit for the data with which it was developed, it performed poorly on the testing data suggesting the initial good fit was the result of chance relations in the data set rather than any significant underlying pattern. As with the majority of models seen so far, this model faces significant challenges in that a significant number of 'satisfactory' reviews would have to be prioritised in order to successfully identify all the 'unsatisfactory' providers and only just over half – 5 out of 8 – 'unsatisfactory' providers were identified in the riskiest half of providers. Therefore, the answer to the question

> *Using naturally-complete metrics, could the aggregated outcome of QAA reviews of teaching and learning at alternative providers have been successfully predicted?*

is no. Regardless of our definition of success the best available model performed poorly.

In seeking to determine which metrics, if any, would have successfully predicted the overall outcome of QAA reviews of *teaching and learning* at alternative providers, this analysis has considered all metrics with a feasible link to quality assurance, even those not centrally available and requiring significant resource to source and process, not just in their absolute state but also

modified to account for changes over time. The best predictive model was still not able to successfully identify a high-risk group of providers who were significantly more likely to be judged 'unsatisfactory' in relation to *teaching and learning* than the low-risk group.

Of the two question-level analyses conducted thus far one has been slightly more accurate than the overall model and the other has been poor. The next stage is to conduct the final question-level analysis.

7.5. Results – The Provision of Information

Using naturally-complete metrics, could the aggregated outcome of QAA reviews of the provision of information at alternative providers have been successfully predicted?

7.5.1. Initial Data Exploration

The outcome of the *provision of information* question overlaps less with the overall review outcomes when compared with the *academic standards* and *teaching and learning* review questions. Table 7.12 below shows that of 41 reviews which were 'unsatisfactory' overall, 22 were 'unsatisfactory' in relation to *the provision of information.*

	Unsatisfactory		Satisfactory	
	Does not meet UK expectations	Requires improvement to meet UK expectations	Meets UK expectations	Commended
Overall review outcome	41		264	
Provision of information	21	1	283	0

Table 7.14: The results for the overall outcome and the provision of information section of each alternative provider review.

Closer inspection reveals that 16 of the 21 providers judged 'unsatisfactory' in relation to *the provision of information* were also judged 'unsatisfactory' in relation to *academic standards.* It is therefore probable that the *provision of information* model and a binary 'satisfactory' / 'unsatisfactory' model for the *academic standards* question would be similar.

Six metrics had a p-value less than 0.25: five finance metrics, two of which were not significant predictors of the overall review outcome, and the provider's age at the time of their review.

Metric Code	Metric Description	P-value
APA009	Investments / Stocks	0.000
APA001	Age at time of review	0.020
APA011	Cash At Bank And In Hand	0.081
APA013	Creditors: Amounts Falling Due Within One Year	0.085
APA002	Outstanding Mortgage Charges	0.159
APA006	Tangible Assets	0.200

Table 7.15: A breakdown of all metrics from the provision of information data set with a p-value of less than 0.25.

As expected given the overlap in question-level outcomes, five of the six significant metrics were also significant in the *academic standards* analysis. The more significant of the two metrics not yet shown in detail – *APA013 Creditors: amount falling due within one year* – is shown in Figure 7.27 below.

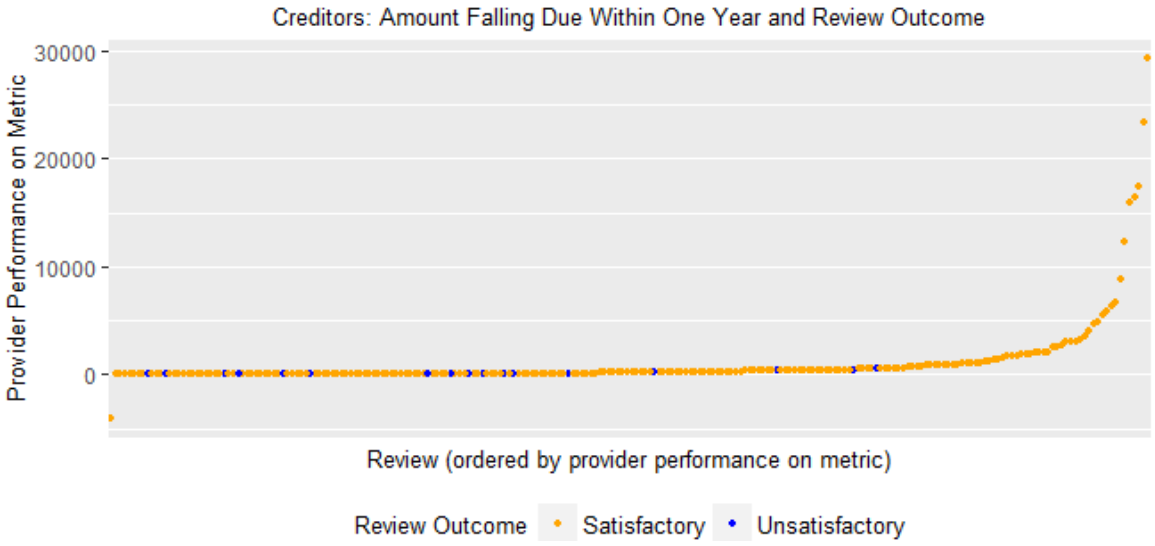

Figure 7.27: The amount each provider has falling due to creditors within one year at the time of their review and the outcome of their review.

The distribution is similar to the *APA001 – Cash at bank and in hand* metric. Moreover, the pattern is now extremely familiar: a subset of 'satisfactory' providers can be identified but the 'unsatisfactory' providers are distributed evenly amongst a large number of 'satisfactory' providers. The metric is therefore of limited use in accurately identifying 'unsatisfactory' provision.

7.5.2. Fitting the Model

The λ_{min} model contains eight metrics, five fewer than the λ_{1se} model which appears to be overfitting the data given the number of metrics and the size of their coefficients.

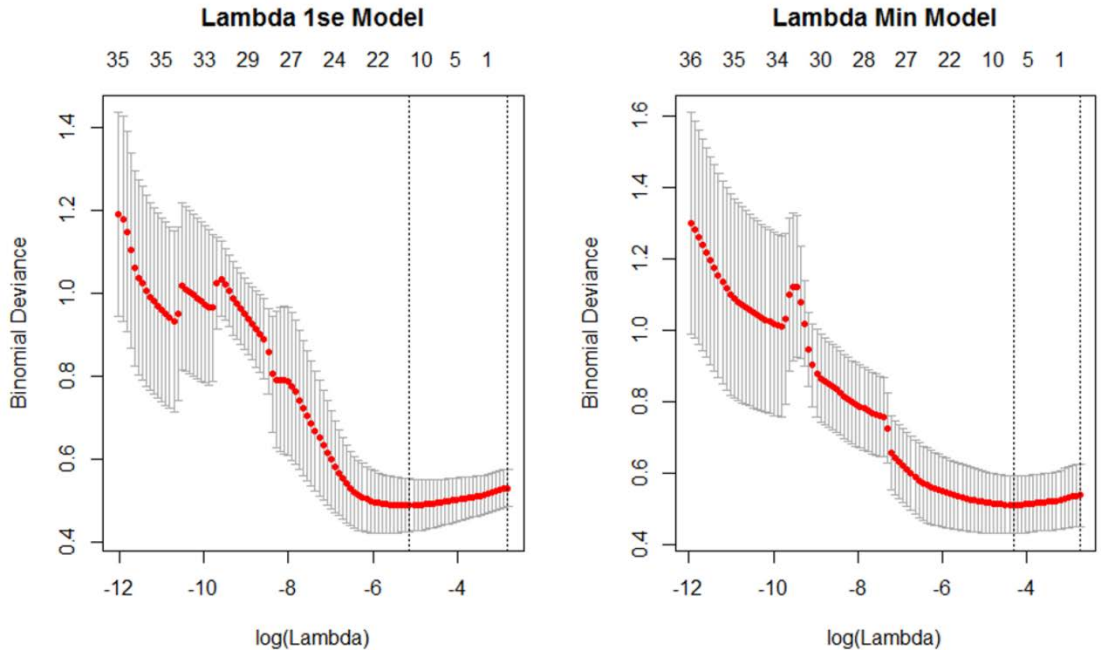

Figure 7.28: The diagnostic plots for the λ_{1se} (left) and λ_{min} (right) model for the provision of information model.

The λ_{min} model contains the following metrics:

- APA004 - Accounts Type
- PRV002 - Has been reviewed and received a negative outcome in the last review
- PRV003 - Has ever received a negative review
- APA001 - Age at time of review
- APA002 - Outstanding Mortgage Charges
- APA011 - Cash At Bank And In Hand
- APA018 - One-year change in total net assets/(liabilities)
- CON002 - Count of *QAA concerns* raised which do not relate to quality, standards, information or enhancement (and are therefore automatically invalid)

Financial ill-health, a lack of experience, and student concerns reported to QAA (albeit concerns not relating to systematic quality assurance failings) could all feasibly lead to an increased likelihood of quality assurance failures. The specific λ_{min} model is similar to the models seen already in this chapter with a small number of additional metrics and calculates the probability of an 'unsatisfactory' review as:

$$P(Unsatisfactory) = \frac{e^A}{1 + e^A}$$

where:

$$A = \text{-2.72} + (\text{-0.00005 x APA001}) + (\text{-0.03 x APA002}) + (\text{-0.65 x APA004.FUL}) + (0.83 \text{ x APA004.TES}) + (\text{-0.00003 x APA011}) + (\text{-0.0001 x APA018\_Ca1}) + (\text{-0.4 x PRV002.YES}) + (\text{-0.38 x PRV003.YES}) + (\text{-0.08 x CON002})$$

and:

APA004.FUL = 1 if the provider is classified as a 'FULL' body by Companies House, 0 otherwise
APA004.TES = 1 if the provider is classified as a 'Total exemption SMALL' body by Companies House, 0 otherwise
PRV002.YES = 1 if the provider been reviewed and received a negative outcome in the last review, 0 otherwise
PRV003.YES = 1 if the provider has ever received a negative review, 0 otherwise

As was the case for the earlier questions in this chapter providers classed as 'Total Exemption SMALL' are more likely to be judged 'unsatisfactory' and providers classed as 'FULL' are less likely to be judged 'unsatisfactory'. The older the provider and the healthier its finances the less likely it is to be found 'unsatisfactory'. As with the previous *teaching and learning* model, those providers who have been the subject of an inappropriate QAA concern are less likely to be 'unsatisfactory'.

7.5.3. Evaluating the Model

7.5.3.1. Testing the Fit of the Model

Figure 7.29 below shows the ROC curve for this model when applied to the data used to develop it. The fairly impressive 'area under the curve' value of 0.797 suggests a reasonable rate of 'unsatisfactory' providers being succesfully prioritised as the threshold criteria for triggering a review is lowered.

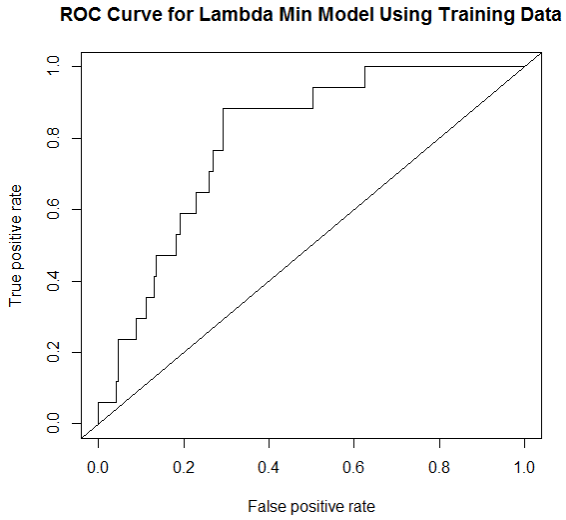

Figure 7.29: The ROC curve for the λ_{min} model fitted to the training data set for the provision of information.

Figure 7.30 and Table 7.14 below show the effect of lowering the threshold required for the model's predicted probability of being judged 'unsatisfactory' to trigger a review.

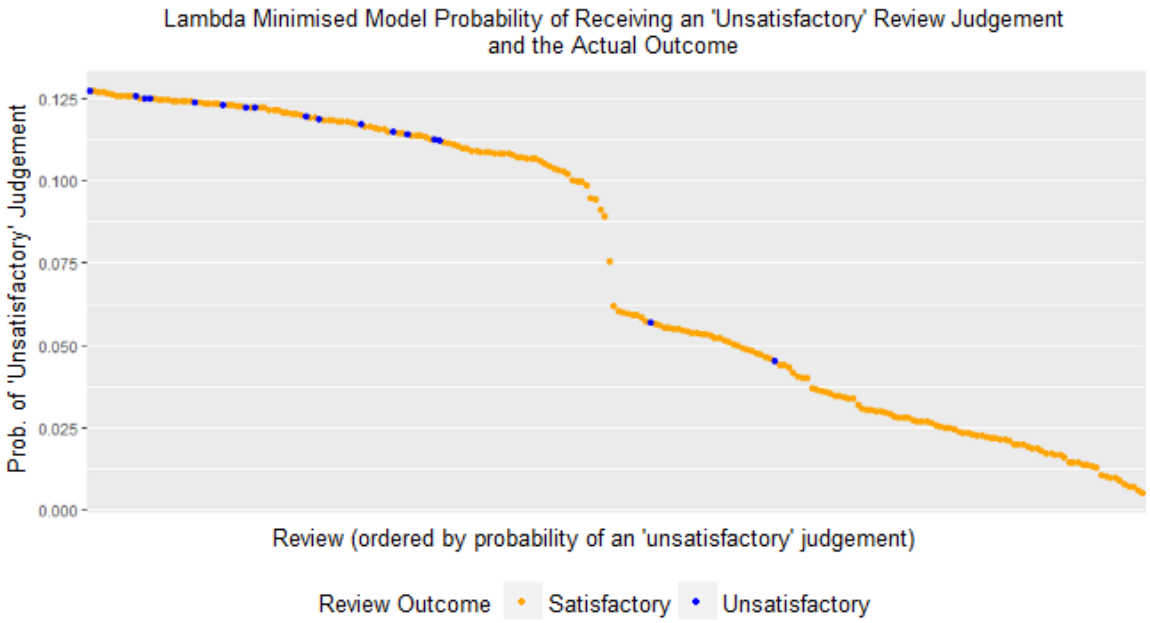

Figure 7.30: Predicted probabilities for each of the 230 complete, comparable provision of information review questions used to train the model and their actual outcome.

	Predicted Probability of 'Unsatisfactory' Review	Number of 'Unsatisfactory' Reviews	Number of 'Satisfactory' Reviews	Accuracy rate
← Decreasing probability of a review being 'unsatisfactory' required to trigger inspection	12.71%	1	0	0.93
	12.48%	4	10	0.90
	12.22%	7	28	0.83
	11.85%	10	41	0.79
	11.39%	13	57	0.73
	5.70%	16	107	0.53
	4.51%	17	133	0.42

Table 7.16: The number of 'satisfactory' and 'unsatisfactory' reviews that would have resulted from decreasing the threshold required to prompt a review concerning the provision of information (only select points are shown).

The model performs reasonably well with all but two of the 17 'unsatisfactory' providers in the training data set are grouped together amongst the higher risk providers. However, the predicted probabilities vary even less than in the earlier models where the range was already concerningly narrow. Moreover, if no 'unsatisfactory' provision is deemed acceptable then, even with perfect hindsight, 133 'satisfactory' providers would be reviewed before all 17 'unsatisfactory' providers were. This represents an error rate of 133 / (133 + 17) = 88.7% amongst those providers prioritised for review: a rate likely unacceptable to providers. Once more it appears what little weak relationship there is between 'unsatisfactory' providers and the metrics also exists for nine times as many 'satisfactory' providers.

7.5.3.2. Assessing the Model's Predictions

Thirty per cent of the reviews of alternative providers which assessed *the provision of information* were held back when developing the model specifically to test its performance. It is apparent from Figure 7.32 below that the model performs well on the testing data. Indeed, with an area under the curve value of 0.829, the model performs marginally better on the withheld testing data than on the training data with which it was developed.

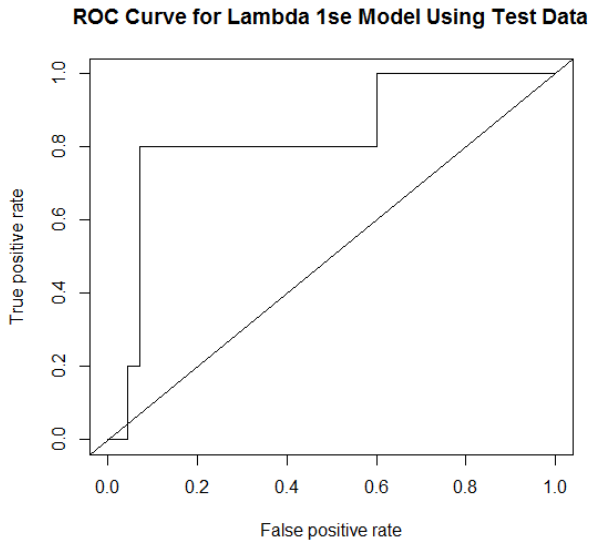

Figure 7.31: The ROC curve for the λ_{min} model fitted to the testing data set for the provision of information.

Figure 7.32 and Table 7.15 below show the model's application to the training set comprising 75 reviews.

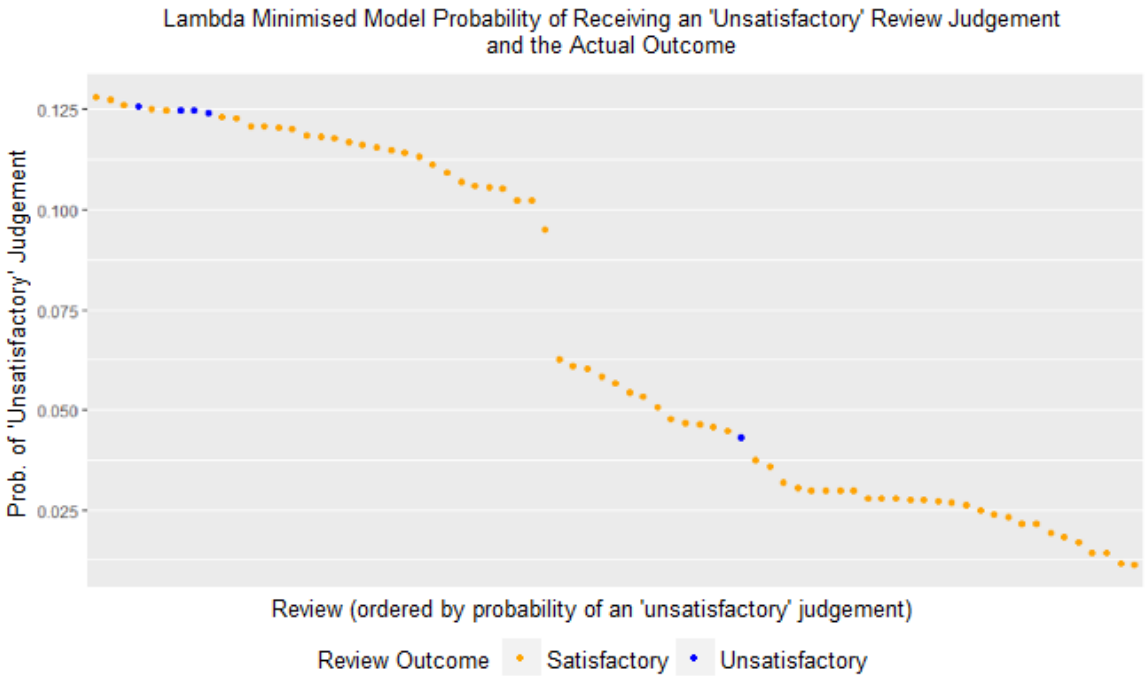

Figure 7.32: The predicted probabilities and actual outcomes of the provision of information review questions contained in the testing set.

	Predicted Probability of 'Unsatisfactory' Review	Number of 'Unsatisfactory' Reviews	Number of 'Satisfactory' Reviews	Accuracy rate
← Decreasing probability of a review being 'unsatisfactory' required to trigger inspection	12.59%	1	3	0.91
	12.49%	2	5	0.89
	12.47%	3	5	0.91
	12.42%	4	5	0.92
	4.31%	5	42	0.44

Table 7.17: The number of 'satisfactory' and 'unsatisfactory' reviews that would have resulted from decreasing the threshold required to prompt a review concerning the provision of information.

The model does a good job of prioritising four out of the five 'unsatisfactory' reviews very early on; however, the final 'unsatisfactory' provider would not have been reviewed until 42 'satisfactory' providers had been incorrectly prioritised. This results in a very high error rate of 89.4%. As with previous models, it appears that the majority, but not all, providers judged 'unsatisfactory' for the *information* question share some similar characteristics; however, so too do many times more 'satisfactory' providers. Further caution must also be expressed as it is not possible to assess the predictions of the model at a specific point in time in order to gauge if the relationships in the model may exist only amongst those alternative providers reviewed. It has also not been possible to assess the models performance against new data to determine whether the relationships which define the model still hold.

7.5.3.3. Summary

The best model calculated the predicted probability of an alternative provider being judged 'unsatisfactory' in relation to their *provision of information* based on previous review performance, the age, size and financial position of the provider, and the number of *QAA concerns* relating to *academic standards* upheld against them. The model determined the predicted likelihood of being judged 'unsatisfactory' as:

$$P(Unsatisfactory) = \frac{e^A}{1 + e^A}$$

where:

$$A = \text{-2.72} + (\text{-0.00005 x APA001}) + (\text{-0.03 x APA002}) + (\text{-0.65 x APA004.FUL}) + (0.83 \text{ x APA004.TES}) + (\text{-0.00003 x APA011}) + (\text{-0.0001 x APA018\_Ca1}) + (\text{-0.4 x PRV002.YES}) + (\text{-0.38 x PRV003.YES}) + (\text{-0.08 x CON002})$$

and:

APA004.FUL = 1 if the provider is classified as a 'FULL' body by Companies House, 0 otherwise
APA004.TES = 1 if the provider is classified as a 'Total exemption SMALL' body by Companies House, 0 otherwise
PRV002.YES = 1 if the provider been reviewed and received a negative outcome in the last review, 0 otherwise
PRV003.YES = 1 if the provider has ever received a negative review, 0 otherwise

Nearly all providers judged 'unsatisfactory' in relation to the *provision of information* question were prioritised amongst the most risky both in the training and testing data sets implying it is based on genuine relations in the data. This relationship was not universal however and a minority of 'unsatisfactory' providers were deemed a low priority. As with the preceding models, there appears to be a weak relationship between the 'unsatisfactory' providers and the metrics, but the characteristics shared by 'unsatisfactory' providers are also shared by many times more 'satisfactory' providers resulting in a high error rate. Therefore, the answer to the question

> *Using naturally-complete metrics, could the aggregated outcome of QAA reviews of the provision of information at alternative providers have been successfully predicted?*

is once more dependent on how success is defined. If it were deemed acceptable that two out of the 17 'unsatisfactory' providers went unreviewed, then, with perfect hindsight, 85 'satisfactory' providers could also have been spared a review. To have reviewed all 17 'unsatisfactory' providers however would have required the QAA to have conducted 150 reviews, 133 of which would have resulted in a 'satisfactory' judgement. These two approaches, even with the benefit of perfect

hindsight, still had very high error rates of 80.5% and 88.7% respectively, neither of which are likely to be accepted by providers.

7.6. Overall Summary

Alternative providers are responsible for a small but growing proportion of UK higher education. The limited data requirements placed on alternative providers means that very little is known about their provision; this poses a challenge to developing a data-driven, risk-based approach. One aspect of the alternative provider data set that is advantageous is the number of reviews, and moreover the number of 'unsatisfactory' reviews, that have taken place. This has allowed for the exploration of whether previous QAA review outcomes could have been predicted at both review and question level and also for testing data to be held back for validating any models.

If the QAA were to make decisions on which providers to review based on the analysis of how likely each provider was to perform on each of the four review questions they would face a number of normative challenges. Each of the hundreds of providers would have a 'risk profile' similar to Table 7.16 below detailing their likelihood of being awarded each outcome:

	Academic standards	Learning opportunities	Information	Enhancement
Commended	N/A	1.2%	31.8%	0.3%
Meets	83.5%	43.8%	49.0%	92.3%
Requires improvement	6.9%	32.6%	5.7%	6.8%
Does not meet	9.6%	22.4%	13.5%	0.6%

Table 7.18: Example probabilities of being judged each possible outcome for each question in a QAA review.

Should a provider which is likely to be 'Does not meet UK expectations' on one question but 'Meets UK expectations' for the remaining three questions be prioritised over another provider likely to be 'Requires improvement to meet UK expectations' for two questions and 'Meets UK expectations' for the remaining two questions? Determining which providers should be prioritised over others inevitably results in aggregating the data. What's more, there is no difference in the burden imposed on providers when being reviewed in relation to one targeted question or comprehensively reviewed on all four questions (HEFCE, 2012, 74). Our first question was therefore:

1. Could the overall outcome of QAA alternative provider reviews have been successfully predicted using metrics?

The best model calculated the predicted probability of an alternative provider being judged 'unsatisfactory' overall based on previous review performance and the age and size of the provider. The model was more promising than we have seen for HEIs or FECs: the predictions based on the held-back testing data reassured us that the model was picking up on genuine underlying patterns in the data and the metrics themselves make intuitive sense. These genuine patterns however were weak with the characteristics shared by most 'unsatisfactory' providers also shared by a far greater number of 'satisfactory' providers. The result was that, as with all previous models, a substantial number of 'satisfactory' reviews would have to be prioritised in order to successfully identify all the 'unsatisfactory' providers. Even with perfect hindsight, over 80% of the reviews undertaken would still have been required to identify all 'unsatisfactory' provision. If we were to accept some 'unsatisfactory' provision not being prioritised however the model did a reasonable job: combining the training and testing data sets 34 out of 41 'unsatisfactory' providers were in the riskiest 50% of providers.

Although there are difficulties in operationalising an approach which assigns multiple probabilities to each provider it could be the case that this approach yields more accurate predictions and should therefore be explored. In the most extreme case we would look to make predictions for each possible outcome of a question (as shown in Table 7.16 above) rather than at the aggregated 'satisfactory'/ 'unsatisfactory' level. Our second question was therefore:

2. Could the exact outcome of QAA reviews of *academic standards* at alternative providers have been successfully predicted using metrics?

The λ_{min} model explored in this section contained 16 metrics covering each of the three key data sources: finance, *QAA concerns* and previous review performance. The model, as with all previous models, had a high error rate but it did appear more promising at identifying all 'unsatisfactory' provision. Indeed, when the reviews were prioritised based on the combined probability of being judged 'Does not meet UK expectations' or 'Requires improvement to meet UK expectations', all such providers were prioritised in the riskiest half of providers. More promisingly the same was true when the model was applied to the test data not used in the development of the model. Whether the model could be considered a success is again dependent on whether success is defined as identifying all 'unsatisfactory' providers with the minimum of unnecessary additional reviews, or less stringently as successfully identifying a high-risk group which contains a greater proportion of 'unsatisfactory' providers than the low-risk group. For both the training and test data all of the 'unsatisfactory' reviews were contained in the 50% of riskiest providers; however, even if the QAA had stopped conducting reviews at the optimal point – immediately after the final

'unsatisfactory' review had taken place - of those providers reviewed approximately 80% would have been judged 'satisfactory' and therefore could be seen as being reviewed unnecessarily.

The granular model showed slightly improved performance compared to the overall review level model but struggled to differentiate between those providers who were judged 'Does not meet meet UK expectations' and 'Requires improvement to meet UK expectations'. Indeed, the model performed best when reviews were prioritised by combining the likelihood of a provider being judged 'Does not meet meet UK expectations' and 'Requires improvement to meet UK expectations' and was only a marginal improvement on a far more straightforward binary 'satisfactory' / 'unsatisfactory' model. The two remaining questions for which there was sufficient data, *teaching and learning* and *the provision of information*, were explored but with the judgements aggregated to 'satisfactory' or 'unsatisfactory'. The remaining two questions were therefore:

3. Could the aggregated outcome of QAA reviews of *teaching and learning* at alternative providers have been successfully predicted using metrics?
4. Could the aggregated outcome of QAA reviews of the provision of information at alternative providers have been successfully predicted using metrics?

For the *teaching and learning* question the best model calculated the predicted probability of a provider being judged 'unsatisfactory' using their previous review performance, the age, size and financial position of the provider, and the number of invalid *QAA concerns* raised against them. Whilst the model fit the data with which it was developed well, it performed very poorly on the held-back testing data suggesting the initial good fit was the result of chance relations in the data set rather than any significant underlying pattern. Regardless of our definition of success the best available model performed poorly. To have identified all of the 'unsatisfactory' providers in the test data nearly all - 83 of the 90 - providers would have had to have been reviewed. Just 5 out of 8 of the 'unsatisfactory' providers were amongst the riskiest 50% of providers.

For *the provision of information,* the best model calculated the predicted probability of a provider being judged 'unsatisfactory' using their previous review performance, the age, size and financial position of the provider, and the number of invalid *QAA concerns*. The model performed similarly with the data with which it was developed and the held-back testing data suggesting that it was based on genuine characteristics. As with previous models however, those characteristics were only shared by the majority of 'unsatisfactory' providers and also by a far greater number of 'satisfactory' providers resulting in a very high error rate for the model.

As noted in section 4.3.4.3 whether or not a model can be deemed a success is subjective. The success of the models can in part be determined by the proportion of 'satisfactory' providers prioritised for review (the error rate), and how many reviews were conducted in total before all 'unsatisfactory' provision had been detected. To varying degrees, the overall review level, *academic standards* and *provision of information* models were able to consistently identify a set of characteristics shared by the majority of 'unsatisfactory' providers. However, these characteristics are also shared by a far greater number of 'satisfactory' providers. The result was that, in each case, the error rate was very high. With such a narrow range of predicted probabilities and high error rate, the alternative provider models would be unlikely to win much support from the providers themselves, a significant number of which would suffer the burden of a review, a burden which their competitors may well be spared, knowing they have been singled out by a prioritisation system that is wrong more than 80% of the time. Further to the additional burden, the provider may also suffer reputational damage after being selected as a high risk provider.

One could argue that the error rate is of little concern if a subset, say 50%, of providers could be identified which contained all 'unsatisfactory' provision. Yes, some providers would be incorrectly prioritised for review, but half of all providers could still be spared a review. Although the 'satisfactory' providers that underwent a review may disagree, reviews do not serve to simply identify poor practice, but also share best practice and help rectify small issues before they grow large. The issue here is that, whilst most 'unsatisfactory' providers (and a greater number of 'satisfactory' providers) share the characteristics which allow the model to identify them, not all do. Just one outlying provider would mean that either the size of the high risk group must become so large few if any 'satisfactory' providers are spared review, or some 'unsatisfactory' provision must be accepted.

Another element to consider when judging the success of a model is how it performs on new data. For alternative providers we were able to test the models' performance on withheld testing data from the same original data set and, notwithstanding the *teaching and learning* model, the patterns identified held true for testing data. This suggests that the characteristics shared by 'unsatisfactory' providers (and a greater number of 'satisfactory' providers) which were the basis of the model, were based on genuine underlying relationships between the metrics and the outcome of the review. There are two reasons why we should remain cautious about these results however. First, these predictions are made with perfect hindsight and so represent the optimal performance the QAA could have ever hoped to have achieved. Due to the resource required to obtain further data we cannot examine the picture the QAA would have been presented with at a given point in time; if we could do so it may be the case that those found 'unsatisfactory' were

amongst the lowest risk providers at that point in time and would therefore not have been reviewed. Second, again due to the resource required to obtain new data, we have not been able to apply the model to more recent data to see if the patterns accounted for in the model still hold. As we saw with FECs, relationships between the data and review outcomes can shift dramatically resulting in a previously-valid model becoming less effective than random selection.

In part the success of the alternative provider models would also depend on how they were utilised. As Table 7.16 above demonstrates, determining the order in which providers should be prioritised for review faced with 15 different probabilities for each provider is challenging. Combining binary question-level models would reduce this to four probabilities, but still poses the same challenge: how should those probabilities be combined? Given that two of the three question-level models only represent a marginal improvement on the overall model whilst one is actually worse, no model could be formed for the *enhancement* question, the fact that there is no real reduction in burden when reviewing a subset of the four questions, and the normative challenges posed, using a single overall review level model seems preferable.

Realistically, to be considered a success despite having a high error rate, a model must at least be robust enough to reassure those providers singled out for review that there won't be other providers that have been spared review that will subsequently be judged 'unsatisfactory'. Moreover, the students who stand to suffer if 'unsatisfactory' provision goes undetected would understandably want to know that QAA were confident there were no issues at their provider if it were to be excused from reviews. None of the alternative provider models appear to be that robust or reassuring. It is hard to see them being accepted either by providers or the students that they serve.

7.7. Discussion

The findings for the alternative provider sector were the most promising yet. A successful model that could underpin a data-driven, risk-based approach to prioritising QAA reviews still could not be found however. Even if a successful model could be found, the relatively low number of students, amount of taxpayer money at stake, and low profile of the providers mean that, with any feasible definition of 'impact' arising from quality assurance failures, alternative providers would pose little risk in comparison to HEIs and FECs.

For the analyses detailed in the last three chapters, this thesis had considered thousands of metrics, not just in their natural format but absolute and percentage change-over-time variants,

standardised data to account for sector wide changes over time and performance relative to peers, and imputed the data to account for missing values. In addition to readily available data, this thesis contains confidential data sourced from the QAA, non-standard HE specific data sets for HE in FE, 600 sets of financial accounts bought and transcribed from Companies House, and has used advanced machine-learning techniques not available even five years ago. This thesis has conclusively shown that no effective model exists that could allow QAA to successfully operate a cost-effective, data-driven, risk-based approach to prioritising individual providers for review.

Were the QAA to adopt a data-driven, risk-based approach, some 'satisfactory' providers would be unfairly burdened with additional reviews, potentially distracting them from delivering their 'satisfactory' provision and unfairly stigmatising them as a high risk provider, whilst some 'unsatisfactory' providers would go without a review for an extended period of time to the detriment of students and the reputation of the UK higher education sector.

The obvious question these findings give rise to is *why is there no meaningful relationship between the metrics and the outcome of QAA reviews?* This question is the subject of the next chapter.

Appendix G – Alternative Provider Metrics

The set of 41 metrics used in this study prior to change-over-time calculations being added.

Area	Metric Code	Metric Description
QAA Concerns	CON001	Count of QAA concerns raised, upheld or otherwise, since previous review
	CON002	Count of QAA concerns raised which do not relate to quality, standards, information or enhancement (and are therefore automatically invalid)
	CON003	Count of QAA concerns upheld since previous review which relate to academic standards
	CON004	Count of QAA concerns upheld since previous review which relate to the quality of learning opportunities
	CON005	Count of QAA concerns upheld since previous review which relate to the provision of information
	CON006	Count of QAA concerns not upheld since previous review which relate to academic standards
	CON007	Count of QAA concerns not upheld since previous review which relate to the quality of learning opportunities
	CON008	Count of QAA concerns upheld since previous review
	CON009	Count of QAA concerns not upheld since previous review which relate to the quality of learning opportunities
Previous Review Findings	PRV001	Outcome of previous review
	PRV002	Has been reviewed and received a negative outcome in the last review
	PRV003	Has ever received a negative review
	PRV004	Outcome of previous comparable review
	PRV005	Has had a comparable review and received a negative outcome in the last comparable review
	PRV006	Worst judgement in previous comparable review
	PRV007	Worst judgement concerning standards in previous comparable review
	PRV008	Outcome concerning standards in previous comparable review
	PRV009	Worst judgement concerning learning in previous comparable review
	PRV010	Outcome concerning learning in previous comparable review
	PRV011	Worst judgement concerning information in previous comparable review
	PRV012	Outcome concerning information in previous comparable review
	PRV013	Worst judgement concerning enhancement in previous comparable review
	PRV014	Outcome concerning enhancement in previous comparable review
Accounts information	APA001	Age at time of review
	APA002	Outstanding Mortgage Charges
	APA003	Satisfied Mortgage Charges
	APA004	Accounts Type
	APA005	Called-Up Share Capital Not Paid
	APA006	Tangible Assets
	APA007	Other Fixed Assets
	APA008	Total Fixed Assets
	APA009	Investments / Stocks
	APA010	Debtors
	APA011	Cash At Bank And In Hand
	APA012	Total Current Assets
	APA013	Creditors: Amounts Falling Due Within One Year
	APA014	Net Current Assets/ (Liabilities)
	APA015	Total Assets Less Current Liabilities
	APA016	Creditors: Amounts Falling Due After One Year
	APA017	Other Long-Term Liabilities
	APA018	Total Net Assets/ (Liabilities)

Table 7.19: The set of 41 metrics used in the alternative provider study prior to change-over-time calculations being added.

8. Why Can't the Available Data Predict the Outcome of QAA Reviews?

Risk-based approaches have come to be seen as a regulatory panacea (Black and Baldwin, 2010). Proponents argue that risk-based regulation allows for targeted and effective oversight, freeing compliant actors from the burden of regulation whilst identifying and resolving non-compliance more quickly and thus reducing the cost of regulation. The QAA, having been in the minority amongst oversight bodies for not being 'risk-based', *Students at the Heart of the System* called for it to adopt "…a genuinely risk-based approach, focusing [its] effort where it will have most impact" and to "explore options in which the frequency – and perhaps need – for a full, scheduled institutional review will depend on an objective assessment of a basket of data" (BIS, 2011, 3.19).

Despite their popularity and high-profile failings, the quantitative aspects of risk-based approaches have faced little scrutiny. The one peer-reviewed assessment to date evaluated the Care Quality Commission's current 'Intelligent Monitoring' tool for prioritising inspections of NHS hospital trusts; even with possibly the most comprehensive set of performance data available to any regulator in the world, the metrics and weightings selected by the CQC proved worse at identifying poorly-performing hospital trusts than random selection (Griffiths *et al.*, 2016). This thesis has provided the first comprehensive, empirical analysis of whether *any* selection of metrics and weightings can form part of a successful data-driven approach to quality assurance in higher education; it cannot.

The obvious question, and focus of this chapter, is *why can't the available data predict the outcome of QAA reviews?* HEIs have a vast quantity of data available: we know about students' applications to universities, who the students are, where they come from and how they progress, the staff that teach them, the research the staff perform when not teaching, the financial health of the establishment in which they do their research, the satisfaction of the students who are taught there, and the employment rates and salaries of those who qualify. Yet despite all this we cannot predict how they will fare in their QAA review. Although many hold an innate distrust for metrics in higher education, and for the ability of a review to get to the heart of quality assurance at a provider, the failure of the former to predict the latter will certainly come as a surprise to those who have championed a data-driven, risk-based approach.

The *prima facie* reason for the failure of the data to predict the outcome of QAA reviews is that they are simply measuring different things. The qualitative indicators contained in the QAA's Quality Code that act as guide for reviewers focus on processes which are difficult, if not impossible, to capture as a quantitative metric; the vast data sets collected turn out not to be 'about' quality, or related quality processes, in any coherent way. This is obvious, and correct, but

also somewhat reductionist or indeed tautological. No-one believes they are directly measuring precisely the same thing; rather, both QAA reviews and the available metrics are seen as proxies for quality. QAA's belief is that by assuring processes, 'quality' will follow (Williams, 2009; Kimber, 2015). Similarly, HEFCE and others believe that 'quality' is indicated by, or directly defined by, outcome measures such as student satisfaction, class sizes, contact hours, and retention (see for example HEFCE, 2016c; BIS, 2016). As both QAA reviews and metrics are seen as proxies for quality, it has been assumed that one could be used to predict the other. For example, low student satisfaction scores may arise because poor quality assurance processes, which will subsequently be judged negatively by the QAA, fail to prevent poor quality provision and do not promote good quality. Conversely, high student satisfaction rates may be seen to arise in part because good quality assurance processes, which will be judged positively by the QAA, lead to good quality outcomes.

This analysis identifies three possible reasons for the inability of the data to predict QAA review outcomes in practice.

Either

1. no subset of the available metrics is a reliable proxy for 'quality' and/or
2. the outcome of QAA reviews aren't reliable proxies for 'quality', hence the former cannot predict the latter.

Or

3. both a subset of the available data and QAA review outcomes are reliable proxies for 'quality'; however, they are measuring different notions of the subjective and hard to define notion of 'quality'.

As discussed in greater detail below, a mixed-methods approach is used to examine these three reasons. This is followed by a discussion of the *prima facie* reason for the failure of metrics to predict the outcome of QAA reviews; the logical assumptions that must hold true for a data-driven, risk-based approach to work; and an exploration of the three possible reasons why such an approach does not work.

8.1. Data and Methods

An empirical epistemology, as adopted for the statistical modelling, emphasises rationality, determinacy and impersonality with knowledge detached from the history and experience of where it was created to seek general predictive conclusions (Usher, 1997). This approach is

appropriate for determining what available data, if any, accurately predicts the risk of quality assurance failings. However, it is not appropriate for exploring why it is not possible to predict quality assurance failings in UK higher education. This exploratory analysis cannot definitively prove why a data-driven, risk-based approach to quality assurance in UK higher education did not work with current or feasible data. It can however ascertain the logical assumptions and preconditions necessary for such an approach to work and the factors identified by key stakeholders that may undermine those. The analysis combines the findings from in-depth interviews with key stakeholders, the analysis of policy documents and academic literature, written and oral evidence from select committee inquiries, and my own experience shadowing a QAA review.

The first data source explored was the written and oral evidence from two key parliamentary select committee inquiries: the 2009 Innovation, Universities and Science Select Committee *Students and Universities* inquiry and the 2015 Business, Innovation and Skills Select Committee *Assessing Quality in Higher Education* inquiry. These inquiries provided invaluable insight into the development of QAA's review process and current thinking respectively. The *Assessing Quality in Higher Education* inquiry provided oral testimony from 27 key stakeholders including the Chief Executives from QAA, HEFCE, OFFA, OIA, the Minister for Universities and Science, the Director of Higher Education at BIS, QAA reviewers and heads of mission groups including GuildHE and the Russell Group. I also gave oral evidence to a closed session of the committee pertaining to the quantitative findings of this thesis and submitted one of 83 pieces of written evidence the inquiry received (Griffiths, 2015). Whilst not always directly addressing the inability of data to predict the outcome of QAA reviews, these hearings provide invaluable insight into the theoretical underpinnings for QAA's approach and the use of data in higher education regulation.

Alongside the select committee evidence, summarised responses to three relevant consultations were considered: HEFCE's 2012 'A Risk-Based Approach to Quality Assurance' consultation, and the Quality Assessment Review Steering Group's 'The Future of Quality Assessment in Higher Education' 'discussion document' and 'Future Approach to Quality Assessment in England, Wales and Northern Ireland' consultation. These consultation outcomes, summarised by HEFCE with the exception of the QARSG's 2015 'discussion document' summarised by MRUK Research, provide a breakdown of responses from interested parties and select quotes.

Semi-structured interviews were then conducted with representatives of all the key stakeholders including past and present senior managers from higher education regulatory bodies, QAA reviewers, student representatives from NUS, senior managers and academics at HEIs, FECs,

alternative providers, and a 'Professional, Statutory or Regulatory Body' (PSRB) with responsibility for academic standards of qualifying degrees for their profession. All interviewees were offered anonymity being referred to only by a broad reference to their role to contextualise any comments drawn from the interview. All interviewees except for Roger Brown, former Chief Executive of the Higher Education Quality Council, took up this offer and have been referred to by a general descriptor.

These interviews were used to explore and develop the themes emerging from the documentary analysis, to gain new insights as to why data cannot predict the outcome of QAA reviews, and to gain an understanding of the reasoning behind certain decisions that the documentary analysis could not provide. The interviews were not designed to solely examine hypotheses developed from the documentary research, but also to examine whether further hypotheses existed and to explore them. Any new ideas generated in the interviews led to further documentary analysis and the reshaping of subsequent interviews to further explore these topics. As such, the research did not use either a purely deductive or purely inductive approach, but an abductive approach which contained elements of both (Peirce, 1958; Locke *et al.*, 2008)

Further to the interviews, I shadowed QAA reviewers on a three-day *Higher Education Review.* This involved observing interviews with senior leaders, students, information specialists and employers working with the institution. Furthermore, I observed the deliberative process of reaching a review judgement and deciding upon commendations and recommendations for improvement. I have also spent considerable time with QAA staff developing an in-depth understanding of the organisation and shaping ideas for the thesis.

As an interviewer and observer it was necessary for me to acknowledge the influence my presence may have (Alasuutari, 1995). Although in many circumstances my status as an academic could qualify me as an 'elite' this was unlikely to be true for this study; those interviewed were senior, successful individuals many of whom have been, or continue to be, academics themselves. This may in itself have proven beneficial as, having been in my position, they may be more minded to help me and support the study. Regardless of how I was viewed, the topic will have impacted upon the type of responses I received. The topics discussed were often highly political in nature, especially given the uncertainty over the future of QAA and HEFCE at the time, and the study has the potential to provide an evidence base which may undermine individual decisions and government policy. It is reasonable to assume interviewees were naturally keen to preserve the reputation of their organisations and their own future resulting in measured and politically-filtered responses. I have also been mindful of the influence my professional background as a health and

social care regulator. Whilst this provides valuable insight into the realities of regulatory practice, it may also serve as a hindrance in framing my thoughts and leading to presumptions based on experiences in another sector that are incorrect or unfounded. I cannot change these circumstances, but by acknowledging the reflexive challenges I can be aware of the impact they will have on the construction of knowledge.

Together, the interviews, select committee evidence, consultation responses, academic literature, and review observation provide a comprehensive evidence base with which to systematically explore why performance and resources data cannot predict the outcome of QAA reviews. The exploration begins with the logical assumptions that must hold to successfully operate a data-driven, risk-based approach to quality assurance in higher education. This is then followed by an in-depth exploration of each assumption and its potential flaws.

8.2. Logical Assumptions for Risk-Based Quality Assurance

The QAA does not assess quality of provision directly: it neither observes teaching nor reviews students' work. Rather, higher education providers theoretically ensure the quality of their own provision by validating all new courses, monitoring courses annually, reviewing courses using nationally agreed reference points every five years, and appointing external examiners to check quality and standards. In their own words, "the QAA checks the checking" (QAA, 2012a).

The checks performed by the QAA concern the 19 'expectations' detailed in the Quality Code (2011b) (for full details see chapter two – Appendix B). It is easy to see from the Quality Code why accusations of QAA only being interested in processes and paperwork arise (see for example Alderman, 2009; Charlton, 2001). Academic standards expectations A2.2 and A3.1 for example state:

> A2.2 "Degree-awarding bodies maintain a definitive record of each programme and qualification that they approve (and of subsequent changes to it) which constitutes the reference point for delivery and assessment of the programme, its monitoring and review, and for the provision of records of study to students and alumni."
>
> (QAA, 2011c, p.21)

> A3.1 "Degree-awarding bodies establish and consistently implement processes for the approval of taught programmes and research degrees that ensure that academic standards are set at a level which meets the UK threshold standard for the

qualification and are in accordance with their own academic frameworks and regulations."

(QAA, 2011c, p.23)

The more detailed 11 expectations relating to 'assessing and enhancing academic quality' are accompanied by 129 qualitative indicators to support reviewers in making their judgement. These indicators are similarly qualitative, subjective and requiring of human judgement:

B1.5 "Higher education providers make use of reference points and expertise from outside the programme in programme design and in their processes for programme development and approval."

(QAA, 2012d, p.12)

B2.8 "Higher education providers determine how decisions and the reasons for those decisions are recorded and conveyed to prospective students."

(QAA, 2012e, p.17)

B5.2 "Higher education providers create and maintain an environment within which students and staff engage in discussions that aim to bring about demonstrable enhancement of the educational experience."

(QAA, 2012f, p.8)

The QAA's expectations and related indicators, which heavily emphasise processes, are therefore very different in nature to the available performance and resource data such as application rates, student satisfaction, or financial health. One could therefore simply ascribe the failure of the data to predict the outcome of QAA reviews to comparing proverbial 'apples and oranges'.

This explanation certainly has merit. It is however, as stated earlier, somewhat reductionist. Policymakers did not believe QAA expectations and available performance measures are exactly the same; rather, both are proxies for, or direct measures of, quality. As Peter Williams, then CEO of QAA testified to the Innovation, Universities Science and Skills Committee in 2009:

"… Process and outcomes are very strongly linked. It is not an accident. It is because things are done that other things happen. Because teachers plan their teaching, then students will learn. Because students are guided in their learning, they will learn. It is that careful, systematic approach which is important …"

(Williams, 2009, Q.342)

Furthermore, the QAA contest the charge that they are focused exclusively on processes regardless of the outcomes that result. As one QAA Senior Manager stated when interviewed for this study:

> "A central part of the process going back to audit, are the questions *'Okay, how do you know this thing? Show us evidence this thing's working effectively.'* And that very quickly leads on to discussions about outcomes and data ... We've always been interested in outcomes; the idea that we would give an institution a gold rating regardless of what the data is saying is ludicrous."

Therefore, QAA contend that their assessments consider both processes and outcomes to provide an assessment of quality. One example might be retention rates insofar as having appropriate processes in place should prevent high dropout rates, whilst not having appropriate processes in place makes high dropout rates more likely. Moreover, where retention rates are poor, this will result in the QAA investigating the processes when conducting their review. As the earlier analysis demonstrated, however, continuation rates do not predict the outcome of QAA reviews.

HEFCE and others also contend that performance data is indicative of, or indeed a direct measure of, quality. As one HEFCE Senior Manager told me:

> "What we'll actually be looking at is what the indicators tell us about the quality of academic experience and, ultimately, the outcomes for students."

BIS and others therefore thought it entirely feasible, although we now know it not to be possible, that the available performance data – one measure of quality – could have predicted the outcome of QAA reviews – another measure of quality. For it to have done so, three requirements would have had to have held true.

First, the available data must provide a reliable proxy for quality. *Second*, QAA reviews must provide a reliable proxy for quality. *Third*, even if both the available data and QAA reviews provide reliable proxies for quality, it must be the same definition of quality. Accordingly, there are three possible overarching and reasons why the available metrics cannot predict the outcome of QAA reviews.

Either

1. no subset of the available metrics is a reliable proxy for 'quality' and/or
2. the outcome of QAA reviews aren't reliable proxies for 'quality', hence the former cannot predict the latter.

Or

3. both a subset of the available data and QAA review outcomes are reliable proxies for 'quality'; however, they are measuring different notions of the subjective, contested notion of 'quality' and hence the data will not be able to predict the outcome of QAA reviews.

Each of these three explanations are explored in detail below.

8.3. Data as a Proxy for 'Quality' in Higher Education

Complaints about the use of metrics in the public sector are commonplace (see for example Smith, 1995; Cave *et al.*, 1997; Freeman, 2002). This analysis explores the possible reasons *why* the available data may not be a reliable proxy for 'quality', and hence cannot predict the outcome of QAA reviews. Interviewees identified three overarching reasons: gaming and data quality, the definition and use of metrics, and the granularity and focus of metrics. Select committee hearings concerning the use of data to determine teaching excellence also identified timeliness as an issue and, whilst excellence and quality are not necessarily interchangeable, in this case timeliness of data is a valid concern for assessing quality and is therefore considered. Each of these four reasons is explored in turn below.

8.3.1. Data Quality and Gaming

"Academics are a clever bunch of people. They will optimise behaviours to achieve the best possible outcomes against the indicators that are being used to measure them."

(Wilsdon, 2015b, Q.113)

In interviewing key actors for this chapter, the 'gaming' of metrics was often cited as key reason for the failure of the data to predict the outcome of QAA reviews. Studies in healthcare and education, similar quasi-markets for public goods, show that the use of high-profile metrics often leads to providers seeking to improve their metric performance whilst not improving the underlying performance the metric was designed to measure (see for example Smith, 1995; Propper and Wilson, 2003; De Bruijn, 2002). Higher education is no different (Huber and Rothstein, 2013; Wilsdon, 2015a).

Interviewees stated that the importance placed on university league tables and the 'Key Information Set' (KIS) meant that high-profile metrics, such as the National Student Survey (NSS) and Destination of Leavers of Higher Education (DLHE) survey, were gamed. Examples ranged from the 'vanilla' to outright cheating. At the lower end of the scale one professor cited the allocation

of nearly the entire student engagement budget for undergraduate students to the weeks in the run up to the NSS in their final year. A greater threat to the validity of the measure was revealed to me by a former lecturer who recalled that, shortly before the NSS became available for students to complete at his former institution,

> "colleagues would remind students that '*the universities ranking and reputation are strongly influenced by the NSS, and the value of your degree and attractiveness to employers were influenced by those rankings and reputation. Please bear this in mind when completing your survey.*'"

At the most extreme end of the scale, the University of Derby was recently revealed to have been cheating the DLHE survey by simply pretending not to have made contact with those students who had failed to continue their education or find employment within six months (Ratcliffe and Adams, 2013).

The planned use of retention, satisfaction and employment metrics to measure 'teaching excellence' as part of the upcoming TEF has elicited prospective plans for similar manipulation of key measures. One QAA staff member recalled hearing that the leaders of a chemistry course containing a sandwich year had said "if we're going to be judged on how much our graduates are earning we're going to encourage them to do a placement with a management consultancy rather than with a chemical lab." Similarly, the Principal of Queen Mary's University London testified to the BIS Select Committee's 'Assessing Quality in Higher Education' inquiry

> "if we were so inclined and if there were a funding system that gave us this incentive, the temptation … would be to say, '*There is a very easy way of achieving 100% retention. You minimise your educational standards to the point where no one fails—job done.*'"

(Gaskell, 2015, Q13)

Interviewees clearly felt that the ability of the data to act as a proxy for quality was severely limited by the manipulation and distortion of key measures to the extent that they were unreliable. Despite the frequency with which gaming was cited as a reason to question the ability of data to serve as a reliable proxy for 'quality', it should be remembered that it is difficult, and often fraudulent, to manipulate the vast majority of the thousands of measures considered as part of this analysis. Those measures include *inter alia* financial health, applications, and continuation rates. Whilst it is clear the ability of one or more key metrics to serve as a proxy for 'quality' will be limited by gaming, and hence possibly predict the outcome of QAA reviews, these metrics

represent a small fraction of the overall data set and the effect of gaming on this comprehensive analysis will have been minimal.

8.3.2. Metric Definition and Usage

"There are so many factors affecting future employment it seems to us difficult if not impossible to make a meaningful linkage to teaching quality."

(BIS Select Committee, 2016b, p.10)

Present in the interviews, academic literature, parliamentary hearings and consultations was concern over how some measures have been designed and misused. Metrics do not need to be gamed to be poor or ambiguous indicators of the performance they were designed to measure. For example, the NSS has been criticised on the grounds that higher education is a post-experience good and students are not in a position to fully judge the education they have received, especially in the stressful second term of the third year leading up to final exams (Baker, 2011). Citing the Financial Services Authority (FSA) Task Force on Past Performance's review of customer satisfaction, King (2011a) demonstrated the weakness of such an approach where consumers often lack the ability to understand the complex good they are consuming. Some measures, therefore, are constrained as proxies for 'quality' by their design and their capture of flawed information.

A concern expressed across all forms of evidence considered for this study was that, whilst metrics often do a good job of measuring that which they have been designed to measure, what they are designed to measure is partly outside of the control of the provider, and therefore a questionable proxy for its quality. The employment destination and earnings of graduates, for example, can be affected by *inter alia* the social capital of students, the regional economy, the subject studied and the vocational nature of some jobs. Likewise, as noted by the Business, Innovation and Skills Committee (2016b), retention figures can be affected by personal circumstances of students entirely separate from provider performance, such as family tragedy or financial difficulty. The change in language in recent years from 'Teaching' to 'Learning and Teaching' reflects that higher education is a two-way process. Whilst lecturers can engage and persuade students, no pedagogy can force a recalcitrant student to learn. Where measures are partly outside of the control of providers, there is a clear constraint for that measure acting as a proxy for their quality, and hence its ability to predict another proxy for quality.

Whether within the control of a provider or not, a metric used inappropriately will be of little benefit. This is the case with a large number of metrics that were originally designed as market

information to promote competition between providers touted as metrics for a risk-based approach to quality assurance (BIS, 2011; HEC, 2013; King, 2014a). As one higher education policy analyst told me:

> "The NSS is a very blunt tool. Student satisfaction is not the same as the quality of an experience, nor is it the same as quality in general. A student can be very satisfied with a three-year course because they could go out drinking every night for three years and still get a 2:1"

Indeed, one study has found no correlation between student satisfaction and subsequent performance on standardised Primary Medical Qualification exams (Lancaster and Fanshawe, 2015).

In many cases the metrics being employed as measures of quality are chosen because they are what is available, and because what should be being measured, can't be (Bevan and Hood, 2006; Gibbs, 2010). This was typified by the much criticised inclusion of a 'telephone expenditure per FTE student' metric in the 1989 'University Management Statistics and Performance Indicators' list (CVCP and UFC, 1989; Cave *et al.*, 1997, p.54-7). As far back as 1985 the CVCP/UGC Working Group stressed "there are few indicators of teaching performance that would enable a systematic external assessment of teaching quality to be made" and later cautioned against "concentrating solely on the measurable to the neglect of the wide range of qualitative factors which are impossible to quantify" (CVCP and UGC, 1987, p.4).

A review of the available metrics suggests that, as with *teaching quality*, there are few metrics enabling systematic external assessment of *academic standards*, the *provision of information*, or *enhancement* either. Metrics of higher education inputs, such as funding and the entry qualifications of new students, and outputs, such as degree result and employment rates, lend themselves to easy quantification and are captured. What students experience during their education, such as the degree to which teaching is valued and the quality of feedback to students, are difficult to quantify and are therefore seldom captured in a standardised, centralised form that lends itself to a data-driven, risk-based approach. As one senior academic stated during their interview:

> "We've been at this quite a while haven't we, don't you think if we actually had any decent metrics of student education someone would have discovered them by now? All that happens is that the metrics used distort the whole process."

The available data may therefore not serve as an effective proxy for quality because there are no measures that come close to assessing the key aspects of quality in higher education which are difficult to quantify. Hence, the QAA have adopted the 129 qualitative indicators detailed in the *Quality Code* that need to be assessed in person.

The diversity and autonomy of the higher education sector also provide a challenge for the use of quantitative metrics. As far back as 1985, the Jarratt Report noted that "objectives and aims of universities are defined only in very broad terms" (CVCP, 1985, 3.30). The lack of specified institutional or sector-wide goals for diverse, not-for-profit universities presented a serious problem for quality evaluation (Bourke, 1986). This issue has become more acute as the higher education sector has significantly expanded and diversified in recent years. Even when institutions share clear goals, evidence shows that entrance scores and acceptance rates which predicted the institutional reputation and student graduation rates in large, research universities were far less effective, or wholly ineffective, in smaller, non-selective universities (Schmitz, 1993). If the data that serve as a proxy for quality vary by provider within a sector, an effective, homogenous subset of those data, something practically necessary for an effective data-driven, risk-based approach to operate, will not be able to act as a proxy for quality.

The ability of metrics to serve as a proxy for 'quality', and hence possibly predict the outcome of QAA reviews, can therefore also be constrained by their inability to quantify certain aspects of the 'quality' of provision, misuse, the impact of factors outside of providers' control, or those differences between providers in a diverse sector.

8.3.3. Granularity and Focus

"When you have key performance indicators such as [the ones in this study] and a complex organisation like a university or a college, you can't measure all the elements that you need to … You may have a dozen [indicators], you may have hundreds, but it's a very small projected space from the entire complexity of the structure you are looking at. And therefore you lose hugely the whole picture as soon as you start looking at KPIs."

Interview with QAA Reviewer

A further limitation of higher education data which may impact on their ability to serve as a proxy for quality, that was highlighted by some interviewees, was the level at which the data are aggregated. All data considered as part of this study, and the overwhelming majority of higher education data, are at provider-level. Despite sharing the same institution-level metrics however, universities and FECs are large, complex, decoupled organisations often with significant variation

between and within departments (Weick, 1976; Perrow, 1999). Indeed, in 2015, the University of Liverpool's first-degree courses in dentistry and ophthalmics had the joint highest and lowest course satisfaction rates in the country respectively (HEFCE, 2016b). Provider-level metrics average out pockets of poor and high quality, diluting any signal in metric performance and making it harder for metrics to identify providers with areas of concern. The differences in performance between departments are themselves inimical to the idea that quality assurance processes should ensure a minimum level of quality across an institution. As one QAA Senior Manager said during their interview:

> "To paraphrase Martin Luther King, a threat to quality somewhere is a threat to quality everywhere … your systems can't be working properly if you are having problems and tolerating problems in a particular area."

Whilst data may mask pockets of poor-quality provision, QAA reviewers may be more likely to detect such areas of concern via the student report submitted prior to the review, or through discussions with students during the review. During the author's shadowing of a *Higher Education Review*, a meeting with students quickly highlighted misleading claims made about one specific HND programme and students' ability to transfer into and 'top up' their qualification at other institutions.

More granular data however is not a simple solution. It is often the case that the point at which the data has been reduced down to a grain that is fine enough to be useful, is where the numbers involved become too small to form any meaningful, robust inference from them. Given the size of many programmes this issue cannot be resolved (Brown, 2007; Gibbs, 2010).

Metric focus, how specific or narrowly-focused a metric is, may also be key. Whilst HEIs have a far greater number of metrics, the majority of these metrics are far more narrowly focused than is the case for alternative providers where the predictive models performed marginally better. It may be the case that having 118 very specific HEI finance metrics obscures the 'quality' signal detected in the 18 more general alternative provider finance metrics.

Therefore, the ability of the data to act as a reliable proxy for quality, and hence possibly predict the outcome of QAA reviews, may also be constrained by the data being aggregated to too high a level or being too narrowly focused on very specific aspects of performance or resources.

8.3.4. Timeliness

"It is important to understand that, when you are looking at metrics and data, you are looking at information, typically, from a couple of years ago."

(Gill, 2015, Q.110)

Although not actively proposed as a reason why data may not serve as a proxy for higher education quality by interviewees, attendees at recent select committee hearings were keen to highlight the challenges posed by the timeliness of data (Horseman, 2015; Hiely-Rayner, 2015). When QAA conduct a review, reviewers will speak to staff and students about their ongoing experiences and review the processes in place at the time of their visit. Higher education data, however, is necessarily retrospective and often dated by the time it is available to QAA and others. As was noted by data experts discussing the challenges of using quantitative measures to assess excellence in higher education, data may be of limited use due to the time taken to gather and disseminate the data. Data relating to student qualifiers for example require exams to be completed, final grades awarded and students cleared for graduation. The data must then be gathered centrally and submitted to HESA in a prescribed format. HESA must then collate, quality assure and amend the data and only once this has been completed can the data be published according to certain standards. This process is necessary but slow.

By the time the data is available to bodies such as the QAA to predict the outcome of reviews and take action, the next cohort of students may have already completed a significant proportion of their studies. It may also be the case that an institution has identified and resolved any issues identified when they first collated their data for HESA many months before it was available to the QAA. This will mean there is no longer an issue at the point they are flagged as a concern. Conversely, issues may have arisen at a provider where there was no problem the year before. This will not then be shown in the data until the following year; too late for the students affected.

This issue is particularly severe for employment data gathered by providers six months after students graduate. The data is, arguably, influenced by all three or four years of undergraduate education meaning that by the time it is published it is at best several years out of date and likely unreflective of current performance. The problem is not limited to centralised higher education data; financial accounts for alternative providers do not need to be filed until nine months after the end of the provider's financial year. This means a provider could be in dire financial straits, or even bankrupt, long before the data can show this, or serve as a proxy for quality.

In summary, there are a number of significant issues that limit the ability of the available data serve a proxy for quality, and hence possibly predict the outcome of QAA reviews. Interviewees, supported by a variety of secondary sources, said that metrics may be: gamed, unable to quantify specific aspects of quality provision, misused, outside of the control of providers, fail to account for diversity, too high-level, and or too narrowly focused. Furthermore, evidence from recent parliamentary hearings suggested that a significant proportion of higher education data may be too out-of-date by the time it is available to act as a proxy for quality. These issues have been discussed in higher education for over four decades, yet no solution has been found.

8.4. QAA Review Outcomes as a Proxy for 'Quality' in Higher Education

For a quasi-regulatory body, QAA is well regarded. The summarised responses to HEFCE's initial quality assessment review noted:

> "The role of the Quality Assurance Agency (QAA) as the single body for monitoring and advising on standards and quality in the HE sector was considered to be greatly beneficial, and many stakeholders considered its comprehensive nature to give UK a reputational advantage."
>
> (MRUK Research, 2015, 100)

Similar acclaim for the QAA has been repeatedly echoed by the sector in the media and parliamentary hearings (see for example Tynan, 2015; Scott, 2015; Cooper, 2016).

Internationally, the QAA is held in equal esteem. A European Association for Quality Assurance in Higher Education (ENQA) review found QAA to be the first quality assurance agency to have met all of the Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG). ENQA's report stated:

> "QAA's overall performance against the standards of the ESG is uniformly high. It is a trustworthy, effective and highly credible agency and a leader in the field. QAA is well-led and well-managed at both Board and Executive levels, with a strong Board, which is both well-informed and constructively challenging. The Panel has been consistently impressed by the calibre and professionalism of all those contributing to the work of QAA in maintaining quality and standards across HE in the UK."
>
> (ENQA, 2013, 95)

QAA's experience and expertise has led to multiple, external quality assurance contracts including reviews of Mauritius' tertiary education institutions and courses and providers under contract from the UK's General Osteopathic Council (QAA, 2014a; QAA and GOsC, 2011).

The respect and admiration for the QAA is, however, far from universal. Some feel the QAA is unnecessary in a market environment with one pro vice chancellor telling me that "if QAA ceased to exist we would still need to be excellent or we'd go out of business." Others feel more strongly that QAA are actively harming higher education quality:

> "The QAA, for those of us who have suffered under its tawdry posturing, is a cancer that gnaws at the core of knowledge, value and freedom in education; its carcinogenic growth is now perhaps the greatest pervasive danger to the function of a university as a surviving institution,"
>
> (Docherty, 2008, p.112)

Whilst few interviewees felt this strongly, the focus and robustness of QAA's reviews was the most oft posited reason for the failure of available data to predict their outcomes. More broadly, this analysis identified four overarching factors which might be limiting the effectiveness of QAA reviews, and hence contributing to their inability to act as a proxy for 'quality' or align with available metrics. The inherent limitations of inspections, the over-reductionist nature of review findings, the ineffectiveness of assessing processes, and the decoupling of quality assurance from frontline practices, are each explored in turn below.

8.4.1. The Limitations of Inspection

"I don't think an institution-wide QAA review is going to pick up on [quality], that's just looking at procedures, [at whether we've] got formal procedures across the whole university?"

Interview with University Lecturer

It was the fundamental inability of QAA to 'get to the heart' of 'quality' at a provider that was posited, alongside over-reductionist findings, as the main reason why data was unable to predict the outcome of QAA reviews. Reviews can achieve much that metrics cannot; no metric will detect the culture of a provider or the nuanced, specific concerns of a subset of students. Reviews are, however, constrained by the inspection frameworks, costs, and the normative limitations of inspectors.

QAA reviews are all conducted in the same manner. Providers are made aware that they will be subject to review months in advance. The first stage is for the provider to provide a written

submission along with accompanying evidence, usually hundreds of documents, to QAA detailing how they meet each of the expectations detailed in the *Quality Code* (QAA, 2011b). This is accompanied by a student submission authored by student representatives at the provider. The review team will assess the submissions and accompanying evidence and request additional evidence they deem necessary. The team will then agree upon the length of the visit, whom they will speak to, and what topics they will cover. Once onsite, any additional discoveries can lead to further requests for evidence and / or meetings with relevant parties. Each reviewer is given lead responsibility for a subset of the 16 'expectations' in the *Quality Code* that are assessed. The final day of the site visit involves no further discussions with staff or students but is dedicated to deliberating and agreeing upon the review judgements.

The greatest concern amongst interviewees was simply that the limited interaction with staff and students provided a shallow impression of the provider and allowed for areas of concern to go undetected. As one PSRB which was recently abandoned a similar higher education review approach told me during an interview:

> "When we did visits and had the same issues that are occurring with QAA … it's very qualitative, essentially a lot of the time you are looking at what the uni provides and they're going to be selective and only provide you things they want you to see so it's not a particularly effective way of quality assurance."

An academic from an alternative provider articulated widely stated criticisms of the reviews as 'tick-box' exercises and dominated by paperwork:

> "When we prepare for QAA one of the main preparation tasks is making sure that the paperwork they are going to look at is ready and complete, and you know what they're going to look at really … It is very much a tick-box exercise: is all the paper there? Do the right people know the right answers? And I'm not sure they are really uncovering what the problems are as a consequence."

With the adoption of the *Higher Education Review* approach in 2013 reviews have had a greater focus on students, including interviewing student representatives and the requirement for a student submission, which have made it far harder for providers to hide areas of poor quality affecting students. Other parties, whilst welcoming the approach, were not convinced about the extent to which speaking to students could identify quality assurance issues. As one university lecturer stated "with the students they talk to … we know the sensible students and we could nominate these students".

One student reviewer noted that such deliberate student selection was easily spotted and accounted for:

> "One review I did [the students] felt cherry-picked, that became obvious in the first five minutes. When they try to look perfect we can see through it. For the most part yes they are the keen beans but they represent students well. They are not backward in going forward if you know what I mean."

Reviewers themselves were not overly concerned with the efficacy of the review methods, but did suggest the uniformity of judgements they were able to reach as a reason why the review outcomes may not serve as a consistent proxy for quality. It was not a case of reviewers believing they were correct and others were wrong, rather that different reviewers had different areas of interest and would focus the review accordingly. As one reviewer succinctly stated during an interview:

> "Reviewers have different characteristics, they have different approaches to things, they have different things that are key and uppermost in their mind. Some people have a particular bee in their bonnet about particular things they want to look at. Others have a different approach … Each of those three or four reviewers will home in on particular elements when they come to do their review … they all take out those elements that are of interest to them as much as they possibly can and come to a conclusion based upon what they're focused in on, so you end up with, it seems to me, and this is what has always worried me about QAA reviews, is I've often felt if you had a different set of reviewers going in they could come out with a different result … Even if you did have a review by a group of 4 reviewers that wrote a report up and gave 10 recommendations and 5 [features of good practice] for that, and then if you got a different team in to come and do exactly the same thing, they'd came up with a different set of recommendations and a different set of [features of good practice], and if you had another team they would come up with a different set yet again, and my point is this: they are all correct. None of them are incorrect, they're all coming up with valid points of view."

The consistency of review judgements was raised by some interviewees that were not reviewers, but it was certainly not a consistent concern. One FEC Senior Manager expressed concerns over the reviewers' failure to adapt to assessing different environments:

> "I've found that when inspectors coming from universities would apply their university experience lock, stock and barrel to the college … the incredible focus on minute taking and record keeping I found some of that unrealistic in FE."

The academic literature suggests that, in part, this lack of consistency can be explained by human biases and is not specific to higher education reviews. A human's ability to attribute a value to an item, output or process is constrained by 'anchoring' or 'arbitrary coherence' (Kahneman, 2011). When we first encounter details, say a university league table ranking or a previous reviewer's rating of a provider, that 'anchor' is imprinted in our minds and serves as a point around which our judgements are made. Furthermore, repeated experiments have shown human judgement is affected by the surroundings in which they are made: coffee is perceived to taste better when served in nicer cups and company's accounts are deemed to be in better health when audited in more lavish offices (Ariely, 2009). Arriving at an institution with impressive facilities and a good reputation, deserved or not, inspectors are already primed to judge it favourably. Conversely, a little-known but excellent provider in modest surroundings may be harshly judged. One former auditor with experience of the QAA reviews put forward an alternative suggestion: that the part-time nature of QAA Reviewers meant their skills would not reach the necessary standard:

> "I'm not being critical of any of the individuals involved, they are all experienced and knowledgeable academics - I know they are - but an audit is a different skill and they are not auditors and I can tell they are not auditors cause I am … [Doing three reviews in two years] you're not going to get up to speed."

In their defence, QAA invest a lot of effort to minimise inconsistencies in judgements. Work is divided in such a way that a review team can easily pick up on a rogue reviewer, there are specific requirements for evidence that can be considered and how the judgements that come from it can be reported, the review findings must then go through a moderation panel with no prior knowledge of the review who are there to provide challenge and ensure consistency across reviews. These processes were strengthened following Southampton University's ultimately successful appeal against an 'unsatisfactory' judgement for reasons which remain confidential (Grove, 2013). The same methods have been adopted by a similar quality regulator, the Care Quality Commission, in response to criticism that judgements were not consistent (Francis, 2013) yet in-depth evaluation showed they have failed to eliminate variation severe enough to undermine their regulatory approach (Boyd *et al.*, 2014). It is doubtful that the steps taken by QAA are sufficient to ensure complete consistency across reviews carried out by numerous different

teams, with different backgrounds, different perceptions of quality and limited chance to hone their auditing skills whilst gainfully employed as full-time academics as their primary employment.

Many parties suggested that the outcome of QAA reviews were not able to provide an accurate and consistent assessment of a provider's quality assurance activity, and hence did not serve as a reasonable proxy for quality. It was suggested that some issues could be hidden from reviewers, that the reviews were too focused on paperwork, that reviewers did not review frequently enough to hone their craft, that different review teams would reach different conclusions about the same provider due to their perceptions of quality, and that numerous biases and heuristics could constrain reviewers' ability to reach a valid judgement. In the face of such challenges, it will always be difficult for QAA reviews to serve as a consistent proxy for 'quality', and hence for the data to predict the outcome of QAA reviews.

8.4.2. Over-Reductionist Findings

"QAA is an institution-level review and there is going to be enormous variation within departments, never mind across departments"

Interview with Russell Group Lecturer

Interviewees were in universal agreement over the significant variation in quality and application of processes within higher education providers. Such variation poses a challenge for reviews that result in judgements of provider-level, not department-level or course-level, quality assurance. QAA are reluctant to provide decisions relating only to specific parts of a provider due to the precedent it may set. As one QAA Manager cautioned "before you know it every institution will want to plead for '*ok we'll live with this negative judgement but can you make it clear it's confined to this one programme*'".

QAA do not aim to assess department-level performance, however it often becomes apparent during a review that that there are specific concerns with one or more departments. How then should a provider with nine excellent departments and one 'unsatisfactory' department be judged? While the QAA strive to ensure their judgements are pragmatic, accounting for the extent to which the problem is confined to particular areas, it remains a fact that a single judgement is required to summarise variable quality assurance practices across a provider. Realistically two options exist: acknowledge that central processes are not working if one department is 'unsatisfactory' and therefore find the provider 'unsatisfactory', or acknowledge that 90% of departments are 'satisfactory' and therefore, on balance, the provider is too. Which occurs in practice is not clear. Just as the over-aggregation of data and the loss of information that results

presents a challenge for the use of data as a reliable proxy for quality, so too do the single, provider-level review judgements.

It is not only reaching a judgement across multiple departments that poses a challenge. As one reviewer noted, it is difficult to aggregate a judgement across multiple assessment categories:

> "It always jarred how the categories assessed were so different. Learning resources was assessing libraries, physical resources. Quality enhancement, that is measuring something very different. How do you realistically measure them on the same scale? How do you combine them fairly?"

The size and complexity of institutions, and the subsequent cost of reviewing each individual department, was one of the contributing factors to QAA's adoption of a provider-level assessment focused on central processes. These centralised processes, it has been argued, should assure quality throughout a provider; however, the interviewees widely acknowledged variation within providers, indeed within departments, suggests this is not happening in practice.

It is unlikely that the problems posed by the aggregation of data and review judgements cancel each other out and result in the same view of quality. Rather, the different explicit and implicit weightings present in metric construction and reviews will result in different views on quality meaning neither serve as a reliable proxy for quality. Hence, the data cannot predict the outcome of QAA reviews.

8.4.3. Processes and Outcomes

The specifics of QAA's process-driven approach were not cited by many as a key issue that may result in QAA reviews being a poor proxy for quality. It was, however, a notable concern amongst the more policy-orientated interviewees. The QAA focus on processes for several reasons. *First*, to focus on outcomes instead would not be permitted by the sector they 'co-regulate'. Higher education providers fiercely defend their academic freedom. Previous attempts to impose a more rigorous quality assessment system have been vociferously attacked with QAA reviewers having been denied access to LSE and ministers successfully lobbied to reduce burden (THE, 2001a, 2001b). *Second*, as mentioned above, QAA simply could not afford to make an effective assessment of outcomes. As Roger Brown, former Chief Executive of HEQC, told me during an interview:

> "Although institutions bang on endlessly about the cost of quality assurance, in reality it's much more cost effective to have a small group of people going in every few years to look at processes than it is to have people endlessly crawling over the outcomes"

Third, a belief in the effectiveness of assessing the providers' processes rather than assessing their outcomes and the intractable challenges that accompany that. This is a view shared by many. In October 2015, 169 academics signed a letter to Madeleine Atkins, Chief Executive of HEFCE, and Professor Julia Goodfellow, President of Universities UK, expressing their concern over the planned use of outcomes measures to assess teaching quality (Jones *et al.*, 2015). Concerns included the fact that student attainment is influenced by a host of factors external to higher education provision including student engagement, social class and prior attainment; the obvious incentives for grade inflation; and the narrow concept of quality which does not account for the promotion of lifelong attributes which are difficult if not impossible to capture.

Not everyone agrees on the process-orientated approach however. For example, in 2009, the Innovation, Universities and Skills Select Committee disagreed with Peter Williams' assertion that processes and outcomes "were very strongly linked" (IUSSC, 2009a, 212) and stated that "in not judging the standards themselves, the QAA is taking an unduly limited view of its potential role" (IUSSC, 2009a, 219). As one academic noted in response to the findings in this thesis:

> "The QAA assessments ignore entirely the quality of what's taught. They give top ratings to courses in pure quackery as long as the right bits of paper can be produced, while ignoring the fact that the unfortunate students are being taught pure nonsense. That fact alone makes QAA reports almost useless, and sometimes actually harmful to quality"
>
> (Colquhoun, 2015)

Whether one has a preference for processes or outcomes however, it is true that the existence of processes assessed by the QAA does not necessarily result in quality provision. This is a fact that may clearly limit the ability of QAA reviews to serve as a proxy for quality, and hence be forecast by data also serving as a proxy for quality. If we consider the example of timely student feedback, a provider may have in place a policy which states all feedback should be provided within two weeks of a submission deadline. The first issue is that, unbeknownst to the central administration, the policy may not be followed. Providing the students presented for interview as part of the review process are selected correctly, and the student submission does not highlight the timeliness of feedback as a key issue, the provider can present the correct paperwork and will receive positive QAA review judgement. Second, even if the policy were followed the tight deadline might only be achieved by staff reducing the quality of their feedback, somewhat defeating the purpose of the policy in the first place. In this circumstance the provider could demonstrate the policy is in place, and being followed, and earn a positive QAA review judgement.

In neither case are the students experiencing quality outcomes, and this may be highlighted by the relevant NSS questions, but not by QAA reviews.

One reviewer raised an alternative issue with the link between processes and outcomes. Outcomes data may highlight where there had been an issue, but if reviewers follow-up those poor outcomes and

> "if we find that the provider has noticed that the results are poor and they've actually acted upon that information and put measures in place to prevent that happening again, then we'd say that team was acting correctly and properly, they've got good processes, and therefore we would rate them highly. They are not rated highly on the fact that the data itself is good or bad, but that the processes are in place so that they can understand what the data is telling them, that they're looking at the data and using that to guide whatever they do. And it's what they do that is the important thing, not what the data tells you or not the outcome, it's the process."

Poor outcomes therefore may be the result of a lack of processes, which should be judged negatively by QAA, but they can also guide QAA to good processes, which should be judged positively by QAA. One could argue that the fact poor outcomes were allowed to happen in the first place indicates that in the past processes have not been effective – they should not have allowed outcomes to fall to concerning levels – but by the time QAA conducts their review the provider has had the chance to transform this negative outcome into a positive.

The disconnect between processes and outcomes works both ways, and can also be affected by the time and actions that elapse between when the two are assessed. The absence of process does not necessarily result in a poor outcome. Academics might not need to be told in writing to provide timely feedback, or to provide feedback in a certain way, for students to receive timely and effective feedback. The lack of such a central process may result in an unfavourable judgement from QAA, whilst students experience, and the data demonstrate, quality outcomes. If the presence, or otherwise, of processes does not necessarily lead to quality provision, the ability of QAA reviews to serve as a proxy for quality, and hence be predicted by the available data, will be constrained.

8.4.4. The Decoupling of Quality Assurance Processes

Multiple academics and regulators expressed concern during their interviews over the disconnect that has developed between quality assessment teams within providers and front-line staff interacting with students. Negative QAA reviews can have a significant impact on a provider;

further to inflicting reputational damage, the right for students to access government loans and for the provider to recruit international students can be curtailed. Accordingly, many providers establish quality assurance units to ensure all the necessary processes are in place and can be demonstrated. The result is that, to varying degrees, quality assurance has become decoupled from the provision of higher education (Power, 1997). This decoupling provides one reason why QAA's review judgements may not serve as a reliable proxy for quality, and why the available data fails to predict them.

During interviews, several references were made to course specification documentation compiled by teaching staff being sent to central quality assurance staff who would review the text and make the necessary changes to please QAA. These textual changes had no impact on what was taught or how it was taught, yet despite this disconnect, they could have a substantive impact on QAA's judgements. This decoupling of quality assurance and frontline provision has not escaped the notice of QAA. As one former QAA Senior Manager acknowledged "QAA's approach has reached maturity, people know how to play the game." In an effort to combat this, QAA reviews have expanded to take representations from all stakeholders including student submissions, as one current QAA Senior Manager stated:

> "You're trying to not just talk to representatives because mini-bureaucracies are created there … but you also want to try and talk to normal students about their experiences, get it from the horse's mouth as it were, and speak to staff as well. Normally review teams will talk to particular kinds of staff about issues they are concerned about. If there's a particular issue about induction then, they'll want to talk to new staff. So the interactions are not all with the quality office or whoever the institutions put forward to represent themselves, you're trying to get behind that as well."

Interviews with academics revealed scepticism around the effectiveness of these developments. As one FEC Senior Manager stated: "When we've been subjected to QAA reviews, particularly the last two … was that there's an awful lot of paper moved around and I'm not sure they're getting to the bottom of quality".

A further limitation that may prevent QAA review judgements serving as a proxy for quality, and hence be predicted by the available data, is therefore that the QAA review process has become decoupled from higher education provision. QAA are simply conducting a 'ritual of verification' by reviewing centralised documentation entirely distinct from the activities, and quality, of providers (Power, 1997).

In summary, there are significant issues with the validity of the outcome of QAA reviews as a proxy for quality. The inherent limitations of reviewers and reviews, reducing the quality assurance practices of often large and varied institutions down to a single judgement, the potential weak link between processes and outcomes, and the decoupling of those processes from provision, may each be limiting the robustness and consistency of review findings, and hence the ability of the available data to predict them.

8.5. Differing Notions of 'Quality' in Higher Education

"Quality in education' is a subject extraordinarily difficult to come to grips with, and full of pitfalls. There is no single final answer to the quality question, and we should not look for it. But the issue cannot be avoided."

(Ball, 1985, p.97)

"Quality: we know what it is, yet we don't know what it is. But that is self-contradictory, for some things are better than others: that is, they have more quality. But when you try and say what the quality is, apart from the things that have it, it all goes 'poof'. There's nothing to talk about. But if you can't say what quality is, how do you know what it is, or how do you know that it even exists? If no-one knows what it is, then for all practical purposes, it doesn't exist at all. But for all practical purposes, it really does exist. What else are the grades based on? Why else would people pay fortunes for some things and throw others in the trash pile? Obviously some things are better than others. But what's the betterness? So around you go, spinning mental wheels, and nowhere finding any place to get traction. What the hell is quality? What is it?"

(Pirsig, 1974, p.184)

The third possible reason why the available metrics cannot predict the outcome of QAA reviews is that, even if they aren't significantly affected by the issues identified earlier and both serve as proxies for quality, they are measuring different conceptions of the incredibly hard to define notion of 'quality'. Perhaps unsurprisingly, interviewees struggled to vocalise what they thought it was. A HEFCE senior manager, higher education policy analyst, a QAA reviewer, and former vice chancellor respectively defined quality as:

> "the [QAA] way of looking at quality is much more about quality management and it's about the processes, but is that actually quality or is quality … what actually matters to

students in terms of the academic experience and probably most importantly student outcomes and those things for all intents and purposes the old QAA reviews weren't really designed to measure?"

"A good quality learning environment is one that provides challenge and stretch that enables you to develop as a whole person rather than just learning academic knowledge - are you learning skills and critical thought and how to engage in wider society and all those things. But also that a high-quality academic environment is an inclusive one … It's not high quality if only a certain type of person is able to succeed … Also important is the ability for individual - both learners and teachers and also institutions, across the spectrum - to take risks and innovate and get things wrong, and fail"

"It's extraordinarily difficult to define what quality is … it's neither a subjective nor an objective phenomenon which could be easily pinned down, but you recognised it when you saw it."

"A good university is one that provides the best education it can for the students of whom it has charge, and extends, develops and matures those students as much as anyone could do given the resources and to an acceptable standard. So it's not just about value added, the standards do have to come in there which is why you would have to have something like QAA reviews, peer-review to provide you both externally and internally with the evidence."

Quality, it seems, can concern *inter alia* outcomes (centred on retention and satisfaction in HEFCE's new operating model), developing the students as a person, inclusivity and adaptability to students, risk taking, value-added, academic standards, and doing the best with resources available. These can, or cannot, be assessed by a QAA-like peer-review process.

The challenge of defining and assessing the quality of public goods is a relatively new phenomenon. As Donabedian states in his landmark paper on the quality of healthcare:

"There was a time, not too long ago, when … the quality of care was considered to be something of a mystery: real, capable of being perceived and appreciated, but not subject to measurement. The very attempt to define and measure quality seemed, then, to denature and belittle it. Now, we may have moved too far in the opposite direction."

(Donabedian, 1988, p.1743)

The perennial debate over 'what is quality in higher education' began in earnest in the 1980s when increasing participation, competition and pressure on resources led to concerns that quality – however it was defined – could suffer. The debate has yet to be resolved (Gibbs, 2010, 2012). Failure to agree on a definition for quality, let alone a measure of it, is perhaps unsurprising given the complexity of the task (Ball, 1985).

Harvey and Green (1993) have identified five broad, distinctive definitions of quality in higher education, each with multiple variants. *First*, quality can be defined as being exceptional. The traditional notion of quality is something exclusive, distinctive and special (Pfeffer and Coote, 1991). There are no defined criteria for this, as the Universities Funding Council (UFC) assumed in the 1989 Research Assessment Exercise "panels would recognise quality when they saw it" (UFC, 1991, p.5). Alternatively, within the 'quality as exceptional' concept, quality can be perceived as excellence by exceeding high standards. More specifically, high standards of inputs and outputs such as bright students, Nobel prizewinning lecturers and modern laboratories will be perceived as high quality.

Second, quality can be defined as perfection or consistency. This may take the form of 'zero defects': always conforming to specifications rather than attaining high standards. Such an approach focuses on quality at all stages of the higher education process ensuring 'faults' do not occur (Peters and Waterman, 1982). A high quality provider under this definition would be one in which every student receives their education as indicated and intended, regardless of how attainable the standards were. Alternatively, quality can be conceived of as the product of a 'quality culture'. Rather than an excessive focus on no single failure occurring, all staff are responsible for quality, identify when lapses occur and work together to prevent a reoccurrence in a constantly improving environment.

Third, quality can be defined as fitness for purpose. Distinct from quality as something exclusive and exceptional, fitness for purpose conceives of quality as specific to individual criteria (Ball, 1985; Reynolds, 1986). Fitness for purpose can be conceived as conforming to customer – or student – specifications. The obvious challenge to this approach is the variety of outcomes students on the same course, let alone at the same provider, seek and the fact that they may not be best placed to specify what education they should receive. An alternative is to define fitness for purpose in terms of a provider's mission (Houston, 2007). A provider that focuses on poorly-qualified students with little social capital and sees 50% through to graduation and in to related employment can be regarded as higher quality than a provider with the brightest students and best facilities but who does little to make them exceptional; a well-built Mini can be of higher

quality than a poorly built Rolls-Royce (Green, 1994). This approach was adopted in the 1991 Higher Education White Paper which established the HEQC and quality assessment units to "safeguard the best of the distinctive missions of individual institutions" (DES, 1991, p.27) and continues, in part, to be adopted by the QAA (QAA, 2013c, p.6).

Fourth, quality can be defined as value for money. In a theoretical perfect market the quality of higher education provision would, theoretically, be reflected by its price. The higher education market is, however, a long way from perfect with an upper price limit from which fees seldom deviate. Despite the lack of differentiation in fees, quality can still be conceived as the type and volume of provision students receive for their predominantly uniform fees. Mirroring the challenges of defining quality, defining value for money requires an agreement of what should be delivered and how that will be assessed. Several senior academics expressed the view during their interviews that it would be easy to increase contact time at little extra cost by mandating all students to attend unnecessary, high-volume seminars.

Fifth, quality can be defined as transformative. One conception of transformation is the 'value added' to a student: how has a higher education provider enhanced the knowledge, skills and abilities of its students? The concept of quality as 'value added' has recently grown in saliency because, in the eyes of Gibbs (2012, p13), it is the only measure of quality that is effected by educational practices rather than inputs. A second conception of quality as transformation is the empowerment of students. A high quality provider enables students to affect their own transformation which may result in enhanced awareness, critical thinking and confidence.

Further to Harvey and Green's (1993) five broad definitions of quality, a *sixth* is apparent in the regulatory world: 'quality as meeting minimum standards'. Foster care, insurance , and health and social care services, for example, are deemed to be of satisfactory quality if they can meet more attainable criteria designed to root out substandard providers (Department for Education, 2011; Lloyd's, 2015; CQC, 2015b). This minimum standards approach is also used by the QAA, in relation to the 'threshold' academic standards designed to ensure awards are set at the right level (QAA, 2014f).

Each conception of quality has its advantages and disadvantages which contributes to the lack of an agreed definition. Thinking of quality as exceptional would be common sense to many: those providers with the brightest students, most revered academics, and best facilities can easily be perceived as the highest quality. This definition however ignores what the provider does with all its advantages. Moreover, it means other providers doing incredible work with limited resources will not be regarded as high quality.

Ensuring every student receives what is promised is an alternative but narrow view of quality that, amongst other things, ignores the specifics of what is being delivered perfectly. Quality as fitness for purpose has been widely used in the UK over the past 25 years and acknowledges that different providers rightly have different missions; not all universities can be like, nor should they aim to be like, Oxbridge. The disadvantage of fitness for purpose as a measure of quality is that it holds different providers to different standards. A provider that has 99.9% retention but is not reaching the exacting target of 100% it has set itself, may be judged of lower quality than a provider with a retention rate of just 60%, albeit with more challenging students.

Quality as value for money simply raises the secondary question of '*how the value is measured?*' and helps little. Proponents argue that quality as 'value added' provides the only realistic way of measuring what a provider does; measuring outputs such as the number of first class degrees or future earnings is of little value as they are well predicted by inputs such as the prior attainment of students (Astin, 1985; Graham and Thompson, 2001). By measuring 'value added' however a decision must be made on where value must be added and may focus attention on these areas at the expense of course content. Moreover, providers with the brightest students are starting from a high point and will struggle to add as much value as a provider starting with less high-performing students. Minimum standards, when correctly enforced, provide a guaranteed baseline of what can be expected, but offer no further information than that. Such minimum standards approaches can disincentivise providers from exceeding expectations as their investment of resource may not be recognised.

Even once a definition of quality has been decided upon, those seeking to measure and assess quality require specific standards against which judgements can be made. What is it that must be fit-for-purpose or exceptional? And how is that defined? What may be fit for the purposes of an 18 year-old undergraduate, may not be fit for the purposes of a mature part-time postgraduate. What may be exceptional in a taught science course, may differ greatly from what is exceptional in an anthropology research environment (Gibbs, 2010). For the QAA, as discussed earlier, the specific standards are detailed as 19 'expectations' concerning *academic standards*, *teaching and learning*, the *provision of information*, and *enhancement*. For HEFCE, quality can be defined by achieving good retention, student satisfaction and employability outcomes (HEFCE, 2016c). Gaining a fixed view on what quality is becomes yet more difficult when these defined standards are not fixed, but subject to ongoing review and revision (QAA, 2012b; RSS, 2016).

Consistently assessing quality in a manner acceptable to all parties is therefore extremely challenging. Quality is neither tangible nor objective, it is not widgets to be counted or measured,

but a subjective and disputed characteristic of a service. It may be the case then that, despite all the potential issues highlighted in earlier sections, metrics and reviews serve as reliable proxies for quality, but different definitions of the contested notion of quality. Metrics will be interpreted by a model in absolute terms. They do not differentiate between a provider with a high standard of students and a worrying retention rate of 75%, and a provider seeking to widen participation to groups that traditionally forego higher education and doing well to retain 75% of them. The QAA reviews however adopt a combination of minimum standards and fitness-for-purpose views of quality. All providers must have effective systems in place to ensure certain outcomes – minimum standards – and the outcomes ensured should be appropriate for the mission of the provider – fitness for purpose. In the example above, whilst both providers may have a process with the aim of maximising continuation rates, and both providers have the same continuation rates, the provider with a high standard of students may be deemed 'unsatisfactory' by the QAA, but the provider seeking widen participation may be deemed 'satisfactory'.

It could therefore be argued that both QAA review outcomes and the available data serve as effective proxies for quality, but quality is multifaceted and in the 'eye of the beholder'. Both proxies for quality may be correct, but they are proxies for different definitions of quality, and hence one cannot predict the other. It should be remembered however that benchmarking HEI metrics to focus on performance comparable with similar institutions – making their output more of an assessment of fitness for purpose rather than absolute – yielded no improvement in the predictive model.

8.6. Summary and Discussion

The intuitive reason for the failure of the data to predict the outcome of QAA reviews is that they are simply measuring different things. Whilst true, this is somewhat over reductionist. Had people believed they were measuring the same thing there would have been no need to establish and maintain HEQC and QAA; metrics would have sufficed. Rather, both QAA reviews and metrics are seen as proxies for quality. As both are proxies for quality, it was assumed one should be able to predict the other.

This chapter has identified three assumptions that must hold true for the available metrics to predict the outcome of QAA reviews: the available metrics must be a reliable proxy for quality; QAA reviews must be a reliable proxy for quality; and the available metrics and QAA reviews must be proxies of the same conception of quality. There are, however, a number of factors which may render each assumption flawed.

The ability of metrics to act as a reliable proxy for quality may be limited by gaming, poor design, misuse, the extent to which performance is outside of the control of providers, sector diversity, a lack of granularity, and out-dated information. Whilst those metrics affected by gaming, poor design, misuse, and impact of external factors are a small fraction of the total, they are the metrics that all interviewees first mentioned when considering links between data and review outcomes. The remaining measures, such as applications, finance, staff and student characteristics, and research, were less affected by gaming, poor design, misuse, and factors outside of a provider's control, but were believed to be less indicative of quality by virtue of what they measured. The timeliness of the data, although not mentioned by interviewees, is a factor that affects all measures. It is highly probable, therefore, that, even if one does not consider gaming of a subset of key metrics to be an issue, challenges inherent in the quantification of performance and resource data mean that the constrained ability of data to serve as a reliable proxy for quality is partly responsible for its inability to predict review outcomes.

The ability of QAA review outcomes to serve as a reliable proxy for quality, and hence possibly be predicted by the data, may be limited by their summation of often large and varied institutions with a single judgement, the fact that having processes in place does not necessarily result in outcomes and *vice versa*, the decoupling of those processes from provision, and the inherent limitations of inspections and inspectors. All of these concerns are valid, but are neither unique to QAA nor easily overcome. Hospitals, prisons and schools can all be large, complex and loosely-coupled organisations that are hard, if not impossible, to summarise in a small number of one-word judgements necessary to communicate findings to lay audiences. Previous attempts to directly observe provision and overcome accusations of 'focusing on paperwork' have had to be abandoned in the face of fierce resistance as they were perceived as overly burdensome and impinging on academic freedom. Even if a comprehensive assessment and scoring approach could be agreed upon, QAA will only ever be able to minimise, not eliminate, inconsistencies in judgement between reviews. The widespread concern amongst interviewees over the challenges QAA face with the consistency and focus of their reviews suggests that it is also highly probable that the constrained ability of review outcomes to serve as a reliable proxy for quality is partly responsible for its inability to be predicted by the data.

Finally, quality is contested. It can be conceived of as exceptional, consistent, fit for purpose, value for money, transformative, or meeting minimum standards. It is multifarious, subjective, and perceptions of quality may change over time. QAA's traditional approach has been to focus on minimum academic standards and the fitness for purpose of providers' activities. Metrics, however, tend to measure in absolute terms. Neither is necessarily wrong, but if they are different,

it is highly unlikely a proxy for one notion of quality will be able to predict a proxy for another notion of quality.

It is not possible to say definitively the extent to which each these three overarching factors contribute to the inability of the data to predict the outcome of QAA reviews as part of a data-driven, risk-based approach, although it seems highly probable that each factor does contribute to some extent. What can be said is that, following a comprehensive review of the available evidence and a wide-ranging set of interviews, there are multiple, compounding factors that provide what is likely to be an insurmountable barrier to the approach envisioned in *Students at the Heart of the System* (BIS, 2011) from being achieved now or in the future.

9. Discussion

This final chapter discusses the findings of this thesis. Specifically, it revisits the conception of the research question, how it has been answered, and what the findings were. This is followed by a discussion of the limitations of the study and areas for further research, the implications of this thesis for the quality assurance of higher education, and the meaning of the findings in the wider context of risk-based regulation outside of higher education.

9.1. Overview of the Study

The origin of this thesis was the 2011 White Paper *Students at the Heart of the System* that called for the QAA to adopt a risk-based approach to prioritising their reviews. Such an approach was to "depend on an objective assessment of a basket of data, monitored continually but at arm's length" (BIS, 2011, 3.19). There was, however, a comprehensive lack of empirical evidence to inform such an approach. It was not known how best, or indeed if it were at all possible, to succesfully schedule QAA's reviews based on the risk of a provider failing their review. This thesis aimed to answer that question.

There are, broadly speaking, three possible ways in which one could use data to prioritise regulatory activity. *First*, simple, rules-based methods place regulatees into prioritisation groups based on a small number of often contextual metrics selected *a priori*. Simple, rules-based approaches are cheap and simple to explain, however, they can be unfairly discriminatory, ignore important data, and fail to prioritise individual providers. *Second*, data-informed approaches present and aggregate numerous, wide-ranging metrics selected *a priori* to inform prioritisation decisions by one or more experts. Whilst data-informed approaches allow regulators to consider, and be seen to consider, a wide range of performance measures and incorporate their tacit knowledge, the advantages are outweighed by the significant epistemic challenges, issues with consistency, and the fact numerous studies have consistently demonstrated expert interpretation to be at best equal to simple models (see for example Dawes, 1979; Grove *et al.*, 2000; Meehl, 1986). *Third*, data-driven approaches use machine-learning techniques to determine the most accurate prioritisation model from the available data. Data-driven models are simpler and cheaper to run, can be continuously monitored, and produce an output free from human biases. It is this data-driven approach that best fit the stated goal of *Students at the Heart of the System* - namely to target QAA reviews via the "objective assessment of a basket of data, monitored continually but at arm's length" – and offered the greatest chance of success (BIS, 2011, 3.19). A *data-driven* approach was therefore explored.

To determine how best QAA reviews could be prioritised using a data-driven, risk-based approach over 1,000 performance and resource measures were gathered. These data included confidential data sourced from QAA, a data set specifically constructed by HEFCE, and a data set painstakingly constructed by the author from 600 sets of financial accounts purchased from Companies House. Where appropriate, these data have had one and two-year absolute and percentage change-over-time variants calculated, been imputed to account for missing values, standardised to remove in-year variation, and benchmarked to account for performance relative to similar providers. The latest version of each item of data prior to each review was then combined with the results of each relevant QAA review, whether it was for an HEI, FEC, or alternative provider.

To identify the optimal model for predicting the outcome of QAA reviews, and hence prioritsing them, various machine-learning approaches were available. The *elastic net* approach, a dynamic blending of *ridge* and *lasso* logistic regession was chosen as it best fulfilled the criteria for use by the QAA: models are accurate, easy to maintain and update, and simpler to interpret than other models meaning that they could be more easily understood by QAA and higher education providers alike. The *elastic net* approach was used to, in effect, try every possible combination of metrics and weightings to determine the best possible model.

This analysis – the first to comprehensively assess the optimal approach to prioritising inspections of quality in any sector - has shown that no effective model exists that could allow QAA to successfully operate a data-driven, risk-based approach to prioritising individual providers for review. The continual monitoring of a basket of metrics, as envisioned in *Students at the Heart of the System*, will not be successful. The finding that a data-driven, risk-based approach cannot work, by extension means it is extremely unlikely any risk-based approach to prioritising individual providers for review will work; evidence from a host of other sectors and settings has shown the expert selection and interpretation of data will perform, at best, equal to data-driven approaches.

It is certain that were the QAA to adopt a data-driven, risk-based approach, the stated benefits of such an approach will not be realised. A subset of 'satisfactory' providers will be unfairly burdened with additional reviews, distracting them from delivering their 'satisfactory' provision, and unfairly stigmatising them as 'high risk'. At the same time, a subset of 'unsatisfactory' providers will go without a review for an extended period of time allowing poor-quality provision to continue to the detriment of students and the reputation of the UK higher education sector as a whole.

Having established that a data-driven, risk-based approach to prioritising QAA reivews cannot work, a secondary, qualitative analysis was conducted to determine *why* this might be the case. Such an analysis cannot conclusively identify a single factor, or factors, with certainty, but can

highlight areas to be addressed. In-depth interviews with key stakeholders were combined with the analysis of policy documents and academic literature, written and oral evidence from select committee inquiries, and the author's experience shadowing a QAA review. For a data-driven, risk-based approach to work, the available metrics must be a robust proxy for quality; QAA reviews must be a robust proxy for quality; and the available metrics and QAA reviews must be proxies of the same conception of quality. However, a number of factors which may render each assumption flawed were identified including *inter alia* gaming, granularity, misuse of data, and timeliness issues; over-reductionist and inconsistent review findings; and the contested, subjective nature of quality in higher education.

This thesis has therefore provided the first empirical analysis of data-driven, risk-based approaches for quality assurance in higher education and has shown that such an approach will not work. By extension, a data-informed approach selecting metrics *a priori* and making use of expert interpretation is also extremely unlikely to be successful. Whilst the findings from this thesis are not, and can never be, absolutely certain, they are extremely comprehensive and show that the changes required to make a data-driven, risk-based approach possible in the future are very unlikely to be realised. This is a significant contribution to the literature on risk-based regulation. Whilst there have been multiple qualitative studies of the political and operational challenges facing its implementation, this is the first quantitative study to analyse the most fundamental aspect of risk-based approaches: whether or not risk can be accurately determined. No effective model exists that could have allowed QAA to successfully prioritise individual providers for review.

9.2. Limitations of the Study and Areas for Further Research

As with nearly all analyses, this study has its limitations and raises further questions. The foremost limitation of this study is that, much like declaring there is no such thing as a black swan, stating that no risk-based approach can successfully predict the outcome of QAA reviews suffers from the problem of induction (Vickers, 2011). No matter how comprehensive the nature of the quantitative analysis, one cannot rule out the possibility that there is a model or approach that can predict the outcome of QAA reviews, but that it has not yet been found. Such a model could have been overlooked in the earlier analysis for two reasons.

First, it could be the case that there exists data that was not considered as part of this analysis that could predict, or contribute to the prediction of, QAA review outcomes. This is highly unlikely. The analysis considered included all metrics with the slightest feasible connection to quality or quality assurance from the HEIDI database in addition to other restricted or specially convened data sets, including financial information for alternative providers that is at present realistically too resource intensive to form part of a cost-effective, risk-based approach. New data collections could allow for the development of a succesful model in the future; however, none of the stakeholders interviewed for this study were able to identify a single, as-yet-to-be-collected data set that would aid the prediction of review outcomes.

Second, a different modelling approach may be more succesful. It is possible a more complex but more powerful support vector machine or neural network – or indeed an as-yet-undiscovered machine learning approach – could identify an effective model. It is again unlikely however. There was a very clear lack of relationship between the available metrics and the outcome of QAA reviews demonstrated by the univariate analyses. No matter how advanced the analytical technique, it cannot identify a set of relationships that do not exist. Even if a more complex machine-learning approach were to be successful, the use of such an approach is questionable. Complex models are difficult to interpret – it would be difficult to explain to a provider why they are being prioritised for review – and more importantly, they are burdensome and expensive to develop and maintain. To quote Wagner on the imposition of the dual quality *audit* and *assessment* regime of the 1990s "there is a danger that the costs of the whole exercise to the system … will exceed the funds affected by the outcome" (Wagner, 1993, 281).

There was also a limitation imposed on the study by the outcome of the reviews. HEIs and FECs, and to some extent alternative providers, are very compliant. The result is that the analyses in this thesis are attempting to predict an infrequent event, something which will always be more challenging than predicting a frequent event. As a nation, having few 'unsatisfactory' providers is a positive, but it will make an effective, data-driven, risk-based approach to quality assurance more challenging to successfully design and operate. Additional reviews have been conducted since the analyses in this study were performed and future analyses based on those additional reviews would complement the findings of this thesis.

Finally, in the previous section I asserted that a data-driven approach, as envisioned in *Students at the Heart of the System*, could not work and that, by extension, given the literature reviewed in section 3.5.2, it is extremely unlikely that a data-informed approach would work either. This is true, but at present one cannot say for certain that a data-informed approach cannot work. One

can only say that it would be 'ground breaking' if it were to defy all previous studies on the performance of expert interpretation of judgement compared to simple statistical models.

Even if one were to set aside the very considerable concerns over the cost of a data-informed approach, the periodic and subjective judgements being in direct opposition to the continuous and objective rationale detailed in *Students at the Heart of the System*, and the overwhelming body of diverse evidence that has assiduously demonstrated expert interpretation to be at best equal to simple models (see for example Meehl, 1954; Beaver, 1966; Libby, 1976; Einhorn and Hogarth, 1978; Goldman *et al.*, 1988; Ashenfelter, 2008; Montier, 2009), there are significant concerns over the *a priori* selection and interpretation of data. If one considers the most frequently touted *a priori* metric which experts would likely consider, the NSS results, this study has shown that there is a significant relationship between only one of the 22 questions and the outcome of QAA reviews, and that relationship shows the marginally better performing institutions are more likely to be judged 'unsatisfactory' by QAA! It is doubtful that any amount of expert interpretation could turn a collection of measures with such a lack of relationship to the outcome into an effective, risk-based approach. As detailed in the following section however, such an approach has been adopted by HEFCE (HEFCE, 2016c), and the adoption of this approach poses an interesting opportunity for future research which could theoretically add significantly to the limited empirical study of data-informed approaches.

Unfortunately, researching the effectiveness of a data-informed approach would not be straightforward. The only robust way to evaluate the approach would be to form two representative groups of providers. The first would be a control group with its reviews prioritised by performance on metrics alone. The second group would have their reviews prioritised by the expert panel. Reviewers would have to be 'blind', i.e. inspectors would not be able to know which group the provider they were reviewing belonged to, and by extension could not know their metric scores which would likely make their grouping apparent, lest they subconsciously bias the result. Given the low rate of non-compliance in the sector this experiment would need to be conducted over many years to obtain a statistically significant result, and the metrics and review method would have to remain consistent over that time. In both cases the feedback will be limited: false negatives may go undetected and the attribution of success to expert interpretation (or otherwise) may be challenging (Kahneman and Klein, 2009; Tetlock and Gardner, 2016). The robust evaluation of such an approach therefore is, for all intents and purposes, impossible and, arguably, by not allowing reviewers to know why they are at a provider and therefore enabling them to overlook issues affecting staff and students, unethical.

There are, therefore, a number of limitations to this study. The first limitation is that it is not possible to say with absolute certainty that no effective model exists, but has not been found by this study. Logically, this could be the case due to a lack of data or due to not using the correct modelling approach. Both are extremely unlikely however. Not only did this study include all available data that could feasibly form part of a cost-effective, risk-based approach, but no interviewees were able to identify additional, as yet to be collected data that may help. Furthermore, if a more complicated model does exist, which is unlikely given the robustness of the *elastic net* approach and the lack of relation between the individual metrics and the review outcomes, it would likely be too complex for regulatory use. The third limitation is that there were a limited number of reviews, and within those a limited number of 'unsatisfactory' reviews, with which to develop the predictive models. This study did however use all the available data, and in no instance was the lack of reviews, 'unsatisfactory' or otherwise, severe enough to call into question the findings. Finally, if one extrapolates from the findings to say that, if a data-driven approach cannot work, a data-informed approach cannot work either, this faces the same problem of induction that claiming no data-driven model can predict the outcome of QAA reviews. Despite the findings of this study, HEFCE have adopted a data-informed approach that in theory, although likely not in practice, allows for the important evaluation of a risk-based approach; something which is all too infrequent. This new approach, and other possibilities for the quality assurance of higher education, are discussed in the following section.

9.3. The Future of Quality Assurance in Higher Education

As noted in chapter two, higher education policy has been in unprecedented flux since the time this thesis was first conceived and, in 2016, HEFCE, HEFCW and DELNI published their revised operating model for quality assessment (HEFCE, 2016c). The operating model confirmed that future quality assurance work would be conducted by themselves and the winners of six competitive tenders. The largest contract, and the contract most relevant to this thesis, concerns conducting quality assurance reviews and has been awarded to QAA (QAA, 2016b).

Established providers will no longer face cyclical reviews. Instead, the national funding councils will conduct a desk-based 'Annual Provider Review' exercise that will "build on established data analysis and assurance arrangements" to determine those providers in need of a quality assurance review (HEFCE, 2016c, 92). New providers are to enter a four-year probationary period that begins with an initial peer review. During the four year 'development' period, new providers will be

subject to the same desk-based 'Annual Provider Review' as 'established' providers. After four years, probationary providers will undergo a peer review and either be judged 'established', or remain in the probationary process until such time as they become 'established'.

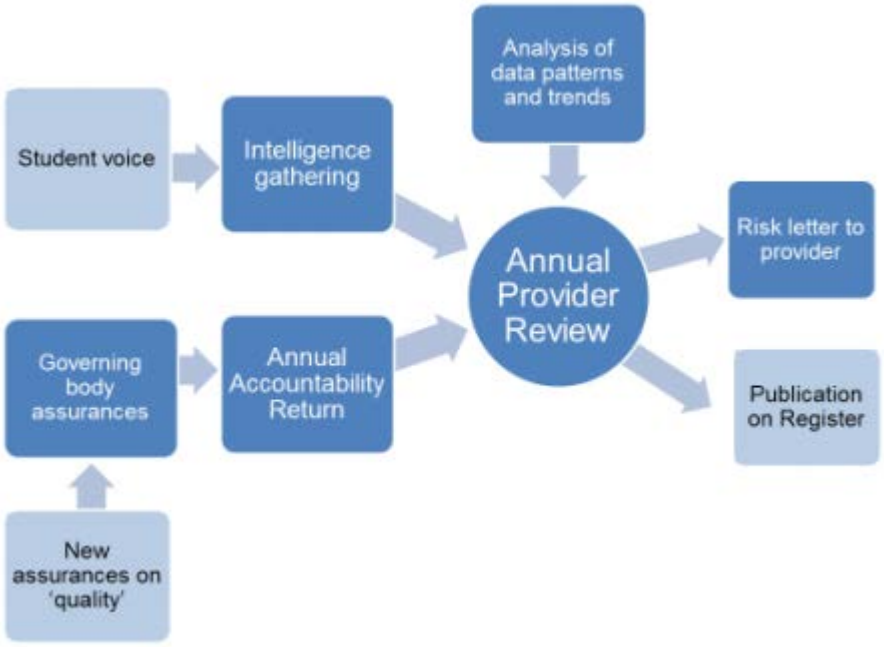

Figure 9.1: Aspects of accountability arrangements in the Annual Provider Review (HEFCE, 2016c, p.24).

HEFCE claim the 'Annual Provider Review' will utilise national funding councils' "sophisticated understanding of its providers and the context in which they operate". Intelligence will be supplemented by the funding bodies "establishing effective ways to capture the views of its students and any outcomes of PSRB activities", and an expanded 'Annual Accountability Return' from governing bodies (HEFCE, 2016c, 94). The 'Annual Accountability Return', which currently requires governing bodies to provide details of their financial sustainability, management, governance, data quality and value for money, will be expanded to include new, unspecified "quality-related assurances" (HEFCE, 2016c, 94). HEFCE have also indicated that, where available, the data used to prioritise reviews will include:

- over- and under-recruitment patterns
- non-progression and non-completion rates
- National Student Survey (NSS) outcomes
- degree outcomes
- employment outcomes
- TEF outcomes

(HEFCE, 2016c, 105).

Typical of risk-based approaches, the specific details of the 'Annual Accountability Reviews' have not been made public. However, HEFCE have noted that data will be "used as one source of information to inform a broader judgement supported where needed by suitably qualified and independent experts" (HEFCE, 2016c, 109). It will be a data-informed approach and a long way from the continual monitoring vision of the 2011 White Paper.

The result of this thesis, and concerns over data-informed approaches, suggest the revised approach is unlikely to successfully differentiate between 'satisfactory' and 'unsatisfactory' providers. It will almost certainly unfairly burden some 'satisfactory' providers whilst allowing 'unsatisfactory' providers to go unnoticed for extended periods. It is worth noting that nearly all the data to be considered as part of the new approach were included in this study. None were convincing metrics of the outcome of QAA reviews.

There is one further option is available for a risk-based approach to quality assurance: a simple, rules-based approach. Analysis of all comparable QAA reviews gathered for this study shows that alternative providers are significantly more likely to be judged 'unsatisfactory' than HEIs ($p=0.029$) or FECs ($p=0.018$). These results were confirmed by QAA at the end of July 2016 when they published analysis of the last two years of reviews showing alternative providers were more likely to be found 'unsatisfactory'. Adding support to the alternative provider analysis in chapter seven of this thesis, the report also confirmed that smaller, younger alternative providers were more likely to have shortcomings (QAA, 2016a). Prioritising alternative providers as a whole over HEIs and FECs may however still be regarded as unfair, will do little to detect 'unsatisfactory' performance in HEIs and FECs, and will generally target the smallest providers in the sector where 'unsatisfactory' provision will arguably have the least impact.

Whatever the reason, or reasons, for the available data not being able to effectively predict the outcome of QAA reviews, it cannot. Therefore, rather than trying and failing to prioritise reviews, an alternative, non-risk-based approach may be advisable. In 2013, recognising the limitations of a risk-based approach to quality assurance in higher education, Brown and Bekhradnia proposed a system of accreditation which must be renewed by all providers on a cyclical basis or immediately following changes of ownership (Brown and Bekhradnia, 2013). Another, non risk-based alternative would be to randomly select providers for review with a guaranteed maximum time elapsing between reviews of each provider. Both non-selective approaches would avoid failed attempts to continually distinguish high-risk providers from others and the costs associated with it. Both approaches would also signal to the providers, and all other stakeholders, that every provider is important. Finally, both approaches would allow all providers to receive the benefits

of the review process; although frequently discussed as a burden, reviews do encourage introspection and enhancement, and allow for the sharing of best practice. Even though the overwhelming majority of reviews result in a 'satisfactory' judgement, all contain a list of recommendations to ensure continued good practice and improvement. Furthermore, with randomised reviews, providers would also need to be incentivised to maintain focus on quality assurance activities knowing that they could be reviewed at any time, rather than losing focus in between cyclical visits. Whatever approach is adopted, there is clear value in empirically evaluating its performance.

A final possible future approach to quality assurance in higher education would be to abolish any form of external oversight. The University of Oxford has existed for over 900 years and nearly all universities predate HEQC and QAA; they managed to survive without external oversight so why should they need it now? As noted in chapter eight, some believe the market in higher education forces providers to maintain the quality of their provision. This argument, whilst it may hint at valid arguments concerning the effectiveness of review methods, overlooks the changed reality of the higher education sector. Immediately following the second world war, less than 2% of further education leavers went on to higher education. Higher education was training for an elite: only 10% of solicitors had higher education and universities had 12 of their own MPs in parliament (Stevens, 2005). Today, nearly half the cohort partake in what is seen as a right for the many, a public service vital for economic health (HESA, 2014). The transformation in higher education attendance has been accompanied by a substantial investment of public funds. Moreover, even if public funds weren't at stake, higher education is a post-experience, positional good. One cannot judge its quality until it is too late, and even if it is poor quality, it may still serve the non-stated purpose of elevating the purchaser's social capital. With more providers, with new motives, in the sector there is more chance that this will be exploited. The market is far from perfect and will not guarantee quality. Whilst the quality assurance system can no doubt be improved, it cannot be done away with.

The revised approach is therefore data-informed and unlikely to successfully identify 'unsatisfactory' providers. Other options are available for a revised approach: simple, rules-based methods or random selection. Doing nothing, however, is not a realistic option.

9.4. Data-Driven, Risk-Based Regulation and the Wider Context

The final question to be addressed is that of how the findings of this thesis can be generalised beyond the quality assurance of higher education. Risk-based approaches to regulation are ubiquitous (Rothstein, 2013). What's more, their use is spreading to all aspects of regulatory activity. As the Government noted in the 2016 higher education white paper:

> "This risk-based approach will not only be at the heart of how the OfS regulates entry to the higher education sector, but in the way it: allows providers access to financial support; assesses and assures ongoing quality; makes judgements on granting degree awarding powers and university title to providers; and ensures that providers have appropriate plans in place to protect students when a provider exits the market"
>
> (BIS, 2016, p.26).

Likewise, CQC are proposing a risk-based approach to registration (CQC, 2015a). No longer are regulators ensuring all providers meet required standards before being allowed entry to a sector, and then inspecting only those providers that they feel are at risk of falling below the defined standards they had previously met. Instead, regulators are now checking only that the riskiest providers meet the required standards before being allowed entry to a sector, and then inspecting only those providers where the regulator feels there is sufficient risk. Given the results of this thesis, this trend is alarming.

The extent to which these findings can be extended to other sectors is not known. The quality assurance of higher education is unique in some regards. The sensibilities around government interference with academic autonomy is unparalleled with the possible exception of the freedom of the press, the regulation of which has been a matter of much unresolved debate and great expense (Leveson, 2012; Greenslade, 2016). Furthermore, QAA is in the minority in not assessing quality itself, but rather that higher education providers have in place the processes to ensure their own quality.

If it is the unique characteristics of QAA that prevents the successful operation of a data-driven, risk-based approach, then such an approach should be able to work for other quality regulators like CQC or Ofsted. Whilst QAA face the challenge of assessing the risks posed by providers that set their own curriculum, exams and mark schemes, primary and secondary education, unlike higher education, has the benefit of a national curriculum and standardised testing allowing pupil performance and progression to be more objectively assessed. One

issue with CQC and Ofsted is that the outcome of their inspections, like QAA, is subjective. Studies have shown a lack of consistency in the judgements of both (Boyd *et al.*, 2014; Fitz-Gibbon and Stephenson-Forster, 2013). Perhaps then a determining factor of whether a risk-based approach will be effective for other regulators is the objectivity of their findings. The Drinking Water Inspectorate is a quality regulator, but their decisions are for the most part objective; levels of hazardous chemicals are either low enough for the water to be safe enough to drink or they are not.

An additional factor that may impact on the success of a data-driven, risk-based approach is the extent to which causal effects can be observed and obey rational, predictable patterns. The most remarkable success of statistical forecasting in the modern era is the accurate prediction of weather up to a week in advance (Silver, 2012). Weather forecasts are based on objective data – wind speeds, pressure, humidity, etc. – and Newtonian interactions of which predictions are readily verified and models updated constantly. The accuracy of a predictive model is clear. Whilst weather forecasting has improved dramatically with advances in computer power, the prediction of earthquakes believed to be caused by unobservable interactions 15km below the earth's surface has not (Hough, 2009). Moreover, economic models continue to disappoint. A lot of economic data can be gathered; however, irrational human behaviour such as spending multiples of one's annual salary on a tulip bulb cannot be predicted (Dash, 2011; Garber, 1990).

In 1814, Simon LaPlace posited that scientific and technological advances may one day bring us to the point whereby all interactions and behaviours are understood, and hence man can foresee the future:

> "We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."
>
> (LaPlace, 1814)

Despite the great advances in computer power – the kind that allows a single research student to consider every permutation of thousands of higher education metrics concerning hundreds of providers over a ten-year period – we have not reached this stage. Moreover, unless we inhabit a world of Calvinistic determinacy, then human irrationality, flawed data and unverifiable models, mean that no matter how much computational power is available, we'll likely never have sufficient foresight to accurately prioritise some regulatory inspections.

What then does this mean for risk-based regulation beyond higher education? The adoption of data-driven, risk-based approaches to regulation has its benefits: it is cheap, objective and empirically-based. However, none of these benefits will be achieved if the data has no relation to the regulatory findings. To implement such an approach regulators must address as best they can the robustness of the available data and regulatory judgements, and how verifiable predictive models will be. Then, any regulator can run a machine-learning exercise to determine if such an approach can be operated successfully. One suspects the chances of success will be higher in environments with meaningful, objective data, robust and verifiable dependent variables, and minimal opportunities for mass irrationality to undermine any model. Such approaches may be more likely to succeed in prioritising financial accounts to review or aeroplanes to safety check say, than which prisons to inspect for quality purposes.

Whether or not a purely data-driven, or data-informed, approach is adopted by a regulator, this thesis has demonstrated the benefit of empirically evaluating such approaches. Without further empirical analysis assiduous platitudes concerning 'making better use of data' or unspecified 'contextualised and nuanced' solutions will simply continue the cycle of regulatory failings followed in quick order by revised 'Intelligence Tools', all whilst people are harmed by regulatory failings potentially more serious than a substandard education. If a regulator cannot demonstrate an effective model, then one must face the politically unpalatable decision of admitting the government, via regulators, cannot protect from all harms, or seek to minimise these potential harms by ceasing the expansion of regulatory responsibilities and reduction of regulatory budgets.

Bibliography

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control,* 19(6), 716-723.

Alderman, G. (1996). Audit, Assessment and Academic Autonomy. *Higher Education Quarterly,* 50(3)**,** 178-192.

Alderman, G. (2009). *Higher Education's Toothless Guard Dog* [Online]. The Guardian. Available: https://www.theguardian.com/commentisfree/2009/may/11/higher-education-select-committee.

AOC (2013). *College Key Facts 2013/14* [Online]. Available: http://www.aoc.co.uk/sites/default/files/AoC%20College%20Key%20Facts%202013-14%20%20web%20format_0.pdf.

AOC (2014). *College Key Facts 2014/15* [Online]. Available: http://www.aoc.co.uk/sites/default/files/AOC%20KEY%20FACTS%202014.pdf.

Ashenfelter, O. (2008). Predicting the Quality and Prices of Bordeaux Wine*. *The Economic Journal,* 118(529)**,** F174-F184.

Astin, A. W. (1985). *Achieving Educational Excellence*, Jossey-Bass.

Atkins, M. (2015). *Letter to Vice-Chancellors and Principals: Quality Assessment in UK Higher Education* [Online]. HEFCE. Available: http://www.hefce.ac.uk/media/hefce/content/news/News/2014/QA/QA_procure_announcement_letter.pdf.

Australian Skills Quality Authority (2016). *Risk-Based Regulation* [Online]. Available: http://www.asqa.gov.au/about/risk-based-regulation/risk-based-regulation.html.

Ayres, I. & Braithwaite, J. (1992). *Responsive Regulation: Transcending the Deregulation Debate*, Oxford University Press.

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic medicine,* 66(3)**,** 411-421.

Baker, S. (2011). *Verdict on QAA Reform: Amateurish and Daft* [Online]. 07/07/2011: THE. Available: http://www.timeshighereducation.co.uk/416733.article.

Baldwin, R., Cave, M. & Lodge, M. (2012). *Understanding Regulation: Theory, Strategy,and Practice*, Oxford University Press.

Ball, C. (1985). *Fitness for Purpose,* Guildford, SRHE & NEFR-Nelson.

Bar Standards Board (2016). *Our Risk-Based Approach* [Online]. Available: https://www.barstandardsboard.org.uk/about-bar-standards-board/how-we-do-it/our-risk-based-approach/.

BBC (2012). *Abortion Clinic Checks Cost £1m* [Online]. Available: http://www.bbc.co.uk/news/health-17620641.

Beaussier, A.-L., Demeritt, D., Griffiths, A. & Rothstein, H. (2016). Accounting for Failure: Risk-Based Regulation and the Problems of Ensuring Healthcare Quality in the NHS. *Health, Risk & Society,* 1-20.

Beaver, W. (1966). Empirical Research and Accounting: Selective Studies. *Chicago, IL: University of Chicago, Graduate School of Business, Institute of Professional Accounting*.

Bendel, R. B. & Afifi, A. A. (1977). Comparison of Stopping Rules in Forward "Stepwise" Regression. *Journal of the American Statistical Association,* 72(357)**,** 46-53.

Berk, R. A. (2008). *Statistical Learning from a Regression Perspective*, Springer Science & Business Media.

BERR (2007). *Regulator's Compliance Code: Statutory Code of Practice for Regulators,* London.

Bevan, G. & Hood, C. (2006). What's Measured Is What Matters: Targets and Gaming in the English Public Health Care System. *Public administration,* 84(3)**,** 517-538.

BIS (2011). *Students at the Heart of the System*. Cmnd 8122. London: BIS.

BIS (2013). *BIS Research Paper No. 111: Privately Funded Providers of Higher Education in the UK* [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207128/bis-13-900-privately-funded-providers-of-higher-education-in-the-UK.pdf.

BIS (2015a). *Alternative Providers of Higher Education: Improving Quality and Value for Money Consultation Document* [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/407730/bis-15-97-alternative-providers-of-higher-education-improving-quality-and-value-for-money-consultation.pdf.

BIS (2015b). *Fulfilling Our Potential: Teaching Excellence, Social Mobility and Student Choice*. Her Majesty's Stationery Office.

BIS (2016). *Success as a Knowledge Economy: Teaching Excellence, Social Mobility and Student Choice* [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/523546/bis-16-265-success-as-a-knowledge-economy-web.pdf.

BIS Select Committee (2015). *Inquiry Launched into Assessing the Quality of Higher Education* [Online]. Available: https://www.parliament.uk/business/committees/committees-a-z/commons-select/business-innovation-and-skills/news-parliament-2015/launch-assessing-the-quality-of-higher-education-inquiry-15-16/.

BIS Select Committee (2016a). *The Teaching Excellence Framework: Assessing Quality in Higher Education*. Third Report of Session 2015–16. London: The Stationery Office Limited.

BIS Select Committee (2016b). *The Teaching Excellence Framework: Assessing Quality in Higher Education*. House of Commons London: The Stationery Office Limited.

Black, J. (2005). The Emergence of Risk-Based Regulation and the New Public Risk Management in the United Kingdom. *Public law,* (3)**,** 512-548.

Black, J. & Baldwin, R. (2010). Really Responsive Risk-Based Regulation. *Law & Policy,* 32(2)**,** 181-213.

Bloch, S. (2014). *Gp 'Disgust' at Watchdog Errors* [Online]. BBC. Available: http://www.bbc.co.uk/news/health-30344455.

Bloom, R. F. & Brundage, E. G. (1947). Predictions of Success in Elementary School for Enlisted Personnel. *Personnel Research and Test Development in the Naval Bureau of Personnel, Princeton University Press, Princeton,* 23361.

Bourke, P. (1986). Quality Measures in Universities. *Belconnen ACT: The Commonwealth Tertiaty Education Commission, Australia*.

Boyd, A., Walshe, K., Robertson, R., Ross, S. & Walshe, K. (2014). Measuring Quality through Inspection: The Validity and Reliability of Inspector Assessments of Acute Hospitals in England. *European Health Policy Group conference.* London:

BRC (2006). *About Us* [Online]. Available: http://www.brc.gov.uk/about.

Brennan, J., Frederiks, M. And Shah, T. (1997). *Improving the Quality of Higher Education: The Impact of Quality Assessment on Institutions*, HEFCE.

Breyer, S. (2009). *Breaking the Vicious Circle: Toward Effective Risk Regulation*, Harvard University Press.

Breyer, S. G., Stewart, R., Sunstein, C. R. & Spitzer, M. L. (1999). *Administrative Law and Regulatory Policy: Problems, Text, and Cases*, Aspen law & business.

Brookman, J. (1992). HEQC Fires Fierce Broadside. 18 Dec ed.: THE

Brown, G. (2005). *Speech to the Confederation of British Industry in London* [Online]. Reported in Financial Times 28 Nov 2005. Available: http://www.ft.com/cms/s/2/9073a120-600d-11da-a3a6-0000779e2340.html#axzz4ILR4cmAW.

Brown, P. R. (1983). Independent Auditor Judgment in the Evaluation of Internal Audit Functions. *Journal of Accounting Research,* 444-455.

Brown, R. (1997). If at First You Don't Succeed … Creating a Single System of External Quality Assurance in UK Higher Education. Institute of Education:

Brown, R. (2004). *Quality Assurance in Higher Education: The UK Experience since 1992,* Oxon, RoutledgeFarmer.

Brown, R. (2007). *The Information Fallacy* [Online]. HEPI. Available: http://www.hepi.ac.uk/wp-content/uploads/2014/03/TheInformationFallacy-RogerBrown.pdf.

Brown, R. & Bekhradnia, B. (2013). *The Future Regulation of Higher Education in England* [Online]. HEPI. Available: http://www.hepi.ac.uk/wp-content/uploads/2014/02/HEPI-Report-63-The-Future-Regulation-of-Higher-Education-in-England.pdf.

Brown, R. & Carasso, H. (2013). *Everything for Sale? The Marketisation of UK Higher Education,* London, Routledge.

BRTF (1997). *Principles of Good Regulation* [Online]. Available: http://webarchive.nationalarchives.gov.uk/20100407162704/http:/archive.cabinetoffice.gov.uk/brc/upload/assets/www.brc.gov.uk/principlesleaflet.pdf.

BRTF (2002). *Higher Education: Easing the Burden,* London, Cabinet Office.

BRTF (2004). *Bridging the Gap: Participation in Social Care Regulation*. London: Cabinet Office.

Cave, M., Hanneyu, S., Henkel, M. & Kogan, M. (1997). *The Use of Performance Indicators in Higher Education: The Challenge of the Quality Movement*, Jessica Kingsley Publishers.

Charlton, B. (2001). *If Not Quality Assurance, Then What?* [Online]. Spiked. Available: http://www.spiked-online.com/newsite/article/11256#.V6mfcJgrKUk.

CHES (1994). *Assessment of the Quality of Higher Education: A Review and an Evaluation,* Institute of Education, University of London.

Clare, J. (2001). University Standards Chief Quits. *The Telegraph*, 22 Aug 2001.

Clark, G. (2015). *Letter from the Rt Hon Greg Clark, Minister for Universities, Sciences and Cities, to Tim Melville-Ross, Chair of HEFCE* [Online]. Available: http://www.hefce.ac.uk/media/HEFCE,2014/Content/Regulation/Course,designation/SNC/Greg_Clark_TimMR_March15.pdf.

Clark, J. (2002). *Non-Prescribed Higher Education: Where Does It Fit?*, ERIC.

Clark, P. (1994). Quality Assessment: Present Position, Future Directions, Address to University of Newcastle Upon Tyne.

Coelho, L. P. & Richert, W. (2015). *Building Machine Learning Systems with Python,* Birmingham, Packt Publishing.

Coglianese, C. & Lehr, D. (2016). Regulating by Robot: Administrative Decision-Making in the Machine-Learning Era. *Georgetown Law Journal,* (forthcoming).

Cohen, A. (1996). Quantitative Risk Assessment and Decisions About Risk. *In:* Hood, C. & Jones, D. (eds.) *Accident and Design: Contemporary Debates in Risk Management.* London: UCL Press.

Colquhoun, D. (2015). *Risk-Based Quality Assessment 'Cannot Work', Study Concludes* [Online]. THE. Available: https://www.timeshighereducation.com/news/risk-based-quality-assessment-cannot-work-study-concludes.

Conservative Party (2015). *The Conservative Party Manifesto 2015* [Online]. Available: https://s3-eu-west-1.amazonaws.com/manifesto2015/ConservativeManifesto2015.pdf.

Cooper, A. (2016). *In Defence of the QAA: The Private Sector View* [Online]. THE. Available: https://www.timeshighereducation.com/blog/defence-qaa-private-sector-view.

CQC (2013a). *A New Start: Responses to Our Consultation on Changes to the Way CQC Regulates, Inspects and Monitors Care Services* [Online]. Available: http://www.cqc.org.uk/sites/default/files/documents/cqc_newstartresponse_2013_14_tagged_sent_to_web.pdf.

CQC (2013b). *Proposed Model for Intelligent Monitoring and Expert Judgement in Acute NHS Trusts (Annex to the Consultation: Changes to the Way CQC Regulates, Inspects and Monitors Care Services)* [Online]. CQC. Available: http://www.cqc.org.uk/sites/default/files/documents/cqc_consultationannex_2013_tagged.pdf.

CQC (2013c). *Quality and Risk Profiles: Statistical Guidance, Outcome-Based Risk Estimates in Qrps Produced for NHS Providers* [Online]. Available: http://www.cqc.org.uk/sites/default/files/documents/20130314_nhs_statistical_guidance_march_2013_for_publication.pdf.

CQC (2014a). *Intelligent Monitoring: NHS Acute Hospitals Frequently Asked Questions* [Online]. Available: http://www.cqc.org.uk/sites/default/files/20141127_surveillance_model_faqs.pdf.

CQC (2014b). *NHS Acute Hospitals: Indicators and Methodology Guidance to Support the December 2014 Intelligent Monitoring Update* [Online]. CQC. Available: http://www.cqc.org.uk/sites/default/files/20141127_intelligent_monitoring_indicators_methodology_v4.pdf.

CQC (2015a). *Building on Strong Foundations* [Online]. CQC. Available: http://www.cqc.org.uk/sites/default/files/20151030_building_strong_foundations_FINAL.pdf.

CQC (2015b). *Guidance for Providers on Meeting the Regulations* [Online]. Available: http://www.cqc.org.uk/sites/default/files/20150324_guidance_providers_meeting_regulations_01.pdf.

Cullen, W. D. (2002). The Ladbroke Grove Rail Inquiry. Sudbury: HSC

CVCP (1985). Report of the Steering Committee for Efficiency Studies in Universities (Jarratt Report). London: CVCP

CVCP (1986). Academic Standards in Universities (Jarratt Report). London: CVCP

CVCP (1991). Published. Cvcp Response to White Paper Higher Education: A New Framework (Cm1541), Letter to DES. July 1991.

CVCP (1992). *Annual Report of the Director 1990/91, Academic Audit Unit,* London, CVCP.

CVCP, CDP & SCOP. 10 Oct 1991. *RE: Quality Assurance Arrangements for Higher Education.* Letter to Kenneth Clark Mp.

CVCP and UFC (1989). *University Management Statistics and Performance Indicators in the UK (Third Edition),* London, CVCP/UFC.

CVCP and UGC (1987). A Second Statement by the Joint Cvcp/Ugc Working Group. London: CVCP

Dash, M. (2011). *Tulipomania: The Story of the World's Most Coveted Flower and the Extraordinary Passions It Aroused*, Hachette UK.

Dawes, R. M. (1979). The Robust Beauty of Improper Linear Models in Decision Making. *American psychologist,* 34(7)**,** 571.

De Bruijn, H. (2002). Performance Measurement in the Public Sector: Strategies to Cope with the Risks of Performance Measurement. *International Journal of Public Sector Management,* 15(7)**,** 578-594.

Deakin, E. B. (1972). A Discriminant Analysis of Predictors of Business Failure. *Journal of accounting research,* 167-179.

Debarr, D. & Harwood, M. (2004). Relational Mining for Compliance Risk. *Internal Revenue Service Research Conference 2004.*

Demeritt, D., Rothstein, H., Beaussier, A.-L. & Howard, M. (2015). Mobilizing Risk: Explaining Policy Transfer in Food and Occupational Safety Regulation in the UK. *Environment and Planning A,* 47(2)**,** 373-391.

Department for Education. (2011). *Fostering Services: National Minimum Standards* [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192705/NMS_Fostering_Services.pdf.

Department for Education (2016). Higher Education and Research Bill. HL Bill 76 56/2

*Deregulation and Contracting out Act, (*1994). Chapter 40, UK. Available: http://www.legislation.gov.uk/ukpga/1992/13/pdfs/ukpga_19920013_en.pdf.

DES (1987). Higher Education: Meeting the Challenge, Cmnd 114. London: HMSO

DES (1991). *Higher Education: A New Framework*. Cmnd 1541. London: HMSO.

Docherty, T. (2008). *The English Question: Or Academic Freedoms*, Sussex Academic Press.

Donabedian, A. (1988). The Quality of Care: How Can It Be Assessed? *Jama,* 260(12)**,** 1743-1748.

Donner, A. (1982). The Relative Effectiveness of Procedures Commonly Used in Multiple Regression Analysis for Dealing with Missing Values. *The American Statistician,* 36(4)**,** 378-381.

Douglas, M. (1992). *Risk and Blame,* London, Routledge.

Dow, K. L. & Braithwaite, V. (2013). *Review of Higher Education Regulation: Report* [Online]. Available: http://www.innovation.gov.au/highereducation/Policy/HEAssuringQuality/Documents/FinalReviewReport.pdf.

Eilers, P. H., Boer, J. M., Van Ommen, G.-J. & Van Houwelingen, H. C. (2001). Published. Classification of Microarray Data with Penalized Logistic Regression. BiOS 2001 The International Symposium on Biomedical Optics, 2001. International Society for Optics and Photonics, 187-198.

Einhorn, H. J. (1972). Expert Measurement and Mechanical Combination. *Organizational behavior and human performance,* 7(1)**,** 86-106.

Einhorn, H. J. & Hogarth, R. M. (1978). Confidence in Judgment: Persistence of the Illusion of Validity. *Psychological review,* 85(5)**,** 395.

ENQA (2013). *Report of the Panel Appointed to Undertake a Review of the UK Quality Assurance Agency (QAA) for the Purposes of Renewal of Full Membership of the European Association for Quality Assurance in Higher Education (Enqa)* [Online]. Available: www.enqa.eu_wp-content_uploads_2014_01_QAA-review-report-FIN2.

Environment Agency (2014). *Operational Risk Appraisal (Opra)* [Online]. GOV.UK. Available: https://www.gov.uk/government/collections/operational-risk-appraisal-opra.

EQAF (2015). *10th European Quality Assurance Forum Hosted by the Quality Assurance Agency and Ucl Institute of Education (19-21 Nov 2015, London)* [Online]. London: EQAF. Available: http://eua.be/Libraries/eqaf-2015/eqaf-2015-programme_04november2015.pdf?sfvrsn=0.

European Court of Justice (2011). Judgement of the Court, Case C-236/09.

Fawcett, T. (2006). An Introduction to Roc Analysis. *Pattern recognition letters,* 27(8)**,** 861-874.

Fawcett, T. & Provost, F. (1997). Adaptive Fraud Detection. *Data mining and knowledge discovery,* 1(3)**,** 291-316.

Field, A., Miles, J. & Field, Z. (2012). *Discovering Statistics Using R,* London, Sage Publications Ltd.

Filskov, S. B. (1984). Clinical Detection of Intellectual Deterioration Associated with Brain Damage. *Journal of Clinical Psychology,* 40(6).

Fitz-Gibbon, C. T. & Stephenson-Forster, N. J. (2013). *Is Ofsted Helpful? An Evaluation Using Social Science Criteria* [Online]. Hoboken: Taylor and Francis. Available: http://kcl.eblib.com/patron/FullRecord.aspx?p=1433905.

Food Standards Agency (2016). *Food Law Code of Practice* [Online]. Available: https://www.food.gov.uk/enforcement/codes-of-practice/food-law-code-of-practice-2015/5-3-frequency-of-controls-and-the-requirements-of-a-risk-based-approach.

Forrest, D., Goddard, J. & Simmons, R. (2005). Odds-Setters as Forecasters: The Case of English Football. *International journal of forecasting,* 21(3)**,** 551-564.

Foster, H. (1993). Right Steps to Standards. *THE*.

Francis, R. (2013). *Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry*, The Stationery Office.

Freeman, T. (2002). Using Performance Indicators to Improve Health Care Quality in the Public Sector: A Review of the Literature. *Health Services Management Research,* 15(2)**,** 126-137.

FSA (2002). Building the New Regulator: Progress Report 2. London: FSA. Feb 2002

FSA(2003). The Firm Risk Assessment Framework. London: FSA. Feb 2003

FSA (2009). *The Turner Review: A Regulatory Response to the Global Banking Crisis,* London, Financial Services Authority.

*Further and Higher Education Act, (*1992). Chapter 13, UK. Available: http://www.legislation.gov.uk/ukpga/1992/13/pdfs/ukpga_19920013_en.pdf.

Garber, P. M. (1990). Famous First Bubbles. *The Journal of Economic Perspectives,* 4(2)**,** 35-54.

Gaskell, S. (2015). Oral Evidence to the BIS Select Committee 'Assessing Quality in Higher Education' Inquiry. London: The Stationary Office Limited.Tuesday 17 November 2015

General Accountability Office (2004). Data Mining: Federal Efforts Cover a Wide Range of Uses (Report No. Gao-04-548) May 2004

Gershon, P. (2004). *Releasing Resources to the Frontline: Independent Review of Public Sector Efficiency*. London: HMSO.

Gibbs, G. (2010). *Dimensions of Quality*, Higher Education Academy York.

Gibbs, G. (2012). Implications of 'Dimensions of Quality' in a Market Environment. *York: Higher Education Academy*.

Gill, J. (2015). Oral Evidence to the BIS Select Committee 'Assessing Quality in Higher Education' Inquiry. London: The Stationary Office Limited. Tuesday 1 December 2015.

Goldberg, L. R. (1965). Diagnosticians Vs. Diagnostic Signs: The Diagnosis of Psychosis Vs. Neurosis from the Mmpi. *Psychological Monographs: General and Applied,* 79(9)**,** 1.

Goldberg, L. R. (1968). Simple Models or Simple Processes? Some Research on Clinical Judgments. *American Psychologist,* 23(7)**,** 483.

Goldman, L., Cook, E. F., Brand, D. A., Lee, T. H., Rouan, G. W., Weisberg, M. C., Acampora, D., Stasiulewicz, C., Walshon, J. & Terranova, G. (1988). A Computer Protocol to Predict Myocardial Infarction in Emergency Department Patients with Chest Pain. *New England Journal of Medicine,* 318(13)**,** 797-803.

Goodhart, C. A. E. (1984). *Monetary Theory and Practice: The UK Experiencie*, Macmillan Publishers Limited.

Gordon, G. (2002). Learning from Quality Assessment. *The Effective Academic: A handbook for enhanced academic practice,* 201-17.

Graham, A. & Thompson, N. (2001). Broken Ranks Us News' College Rankings Measure Everything but What Matters. And Most Universities Don't Seem to Mind. *Washington Monthly,* 33(9)**,** 9-14.

Green, D. (1994). *What Is Quality in Higher Education?*, ERIC.

Greenslade, R. (2016). Press Regulation Battles Fought without Much Public, or Journalistic, Interest. *Guardian*, 23 Aug 2016.

Griffiths, A. (2012). *In Theory & in Practice: Can Risk-Based Approaches to Regulation Work in the Care Quality Domain?* MSc Risk Analysis, King's College London.

Griffiths, A. (2015). *Written Evidence from Alex Griffiths to BIS Select Committee Inquiry into 'Assessing Quality in Higher Education'* [Online]. Available: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/business-innovation-and-skills-committee/assessing-quality-in-higher-education/written/23795.pdf.

Griffiths, A., Beaussier, A.-L., Demeritt, D. & Rothstein, H. (2016). Intelligent Monitoring? Assessing the Ability of the Care Quality Commission's Statistical Surveillance Tool to Predict Quality and Prioritise NHS Hospital Inspections. *BMJ Quality & Safety*.

Grove, J. (2013). *Southampton Shows Teeth and Watchdog Backs Down* [Online]. THE. Available: https://www.timeshighereducation.com/news/southampton-shows-teeth-and-watchdog-backs-down/2003862.article.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. (2000). Clinical Versus Mechanical Prediction: A Meta-Analysis. *Psychological assessment,* 12(1)**,** 19.

Gutierrez, D. D. (2015). *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R,* Basking Ridge, NJ, Technics Publications.

Hampton, P. (2005). *Reducing Administrative Burdens,* London, HM Treasury.

Harrell, F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer.

Harris, R. (1990). The Cnaa Accreditation and Quality Assurance. *Higher Education Review,* 22(3)**,** 34.

Harrison, M., Lockwood, B., Miller, M., Oswald, A., Stewart, M. & Walker, I. (2001). Higher Education: Trial by Ordeal. *Guardian*, 30 Jan.

Harvey, L. & Green, D. (1993). Defining Quality. *Assessment & Evaluation in Higher Education,* 18(1)**,** 9-34.

Hastie, T., Tibshirani, R. & Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* New York, NY, Springer.

Havergal, C. (2015). *Risk-Based Quality Assessment 'Cannot Work', Study Concludes* [Online]. Times Higher Education. Available: https://www.timeshighereducation.com/news/risk-based-quality-assessment-cannot-work-study-concludes.

Hawkins, D. M., Basak, S. C. & Mills, D. (2003). Assessing Model Fit by Cross-Validation. *Journal of chemical information and computer sciences,* 43(2)**,** 579-586.

Healthcare Commission (2009). *Investigation into Mid Staffordshire NHS*

HEC (2013). *Regulating Higher Education* [Online]. Available: http://www.policyconnect.org.uk/hec/sites/site_hec/files/report/333/fieldreportdownload/hecommission-regulatinghighereducation.pdf.

HEFCE (1995). *Report on Quality Assessment 1992-1995,* Bristol, HEFCE.

HEFCE (2012a). *A Risk-Based Approach to Quality Assurance: Consultation* [Online]. HEFCE. Available: A risk-based approach to quality assurance: Consultation.

HEFCE (2012b). *A Risk-Based Approach to Quality Assurance: Outcomes of Consultation and Next Steps* [Online]. HEFCE. Available: http://www.hefce.ac.uk/media/hefce/content/pubs/2012/201227/Risk-based%20quality%20assurance%20consultation%20outcomes.pdf.

HEFCE 24/03/2015 (2015a). *RE: College-Level Higher Education Data.* Personal communcation with A Griffiths.

HEFCE (2015b). *Register of HE Providers* [Online]. Available: http://www.hefce.ac.uk/reg/register/.

HEFCE (2015c). *Register of HE Providers - Overview* [Online]. Available: http://www.hefce.ac.uk/reg/register/search/Overview.

HEFCE (2016a). *How We Fund HE in Fe* [Online]. Available: http://www.hefce.ac.uk/workprovide/hefe/fund/.

HEFCE (2016b). *National Student Survey Results 2015* [Online]. Available: http://www.hefce.ac.uk/lt/nss/results/2015/.

HEFCE (2016c). *Revised Operating Model for Quality Assessment* [Online]. HEFCE. Available: http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/2016/201603/HEFCE2016_03.pdf.

HEQC (1996). Stengthening External Examining. Gloucester: HEQC

HESA (2012). *Students, Qualifiers and Staff Data Tables: Academic Yeare 2011/12 - Student by HE Provider* [Online]. Cheltenham: HESA. Available: https://www.hesa.ac.uk/dox/dataTables/studentsAndQualifiers/download/institution1112.xls.

HESA (2013). *General Student Numbers* [Online]. Available: http://www.hesa.ac.uk/content/view/1897/239/.

HESA (2014). *Free Online Statistics - Students & Qualifiers* [Online]. Available: http://www.hesa.ac.uk/content/view/1897/239/.

HESA (2015). *2013/14 Students by HE Provider* [Online]. Available: https://www.hesa.ac.uk/dox/dataTables/studentsAndQualifiers/download/Institution1314.xlsx.

HESA (2016). *Data Definitions - Estates Management Statistics 2009/10* [Online]. Available: https://events.hesa.ac.uk/index.php?option=com_content&view=article&id=1871&ItemId=233#q70.

Hiely-Rayner, M. (2015). Oral Evidence to the BIS Select Committee 'Assessing Quality in Higher Education' Inquiry. London: The Stationary Office Limited.Tuesday 1 December 2015

Hillman, N. (2014). *A Guide to the Removal of Student Number Controls,* Oxford, Oxuniprint.

HL Deb. §1117 ed.16 December 1991.

HL Deb. §1467-84 ed.21 March 2001.

Hood, C. (1995). The "New Public Management" in the 1980s: Variations on a Theme. *Accounting, organizations and society,* 20(2)**,** 93-109.

Hood, C., James, O., Scott, C., Jones, G. W. & Travers, T. (1999). *Regulation inside Government: Waste Watchers, Quality Police, and Sleaze-Busters*, Oxford University Press.

Hope, C. (2011). £9,000 Tuition Fees Will Be the Exception, Promises Higher Education Minister. *The Telegraph*, 21/02/2011.

Horseman, N. (2015). Oral Evidence to the BIS Select Committee 'Assessing Quality in Higher Education' Inquiry. London: The Stationary Office Limited.Tuesday 1 December 2015

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied Logistic Regression*, John Wiley & Sons.

Hough, S. (2009). *Predicting the Unpredictable: The Tumultuous Science of Earthquake Prediction*, Princeton University Press.

Houston, D. (2007). Tqm and Higher Education: A Critical Systems Perspective on Fitness for Purpose. *Quality in Higher Education,* 13(1)**,** 3-17.

Huber, M. & Rothstein, H. (2013). The Risk Organisation: Or How Organisations Reconcile Themselves to Failure. *Journal of Risk Research,* 16(6)**,** 651-675.

Hutter, B. M. (2005). *The Attractions of Risk-Based Regulation: Accounting for the Emergence of Risk Ideas in Regulation* [Online]. Available: http://www.lse.ac.uk/collections/CARR/pdf/Disspaper33.pdf.

Inwald, R. E. (1988). Five-Year Follow-up Study of Departmental Terminations as Predicted by 16 Preemployment Psychological Indicators. *Journal of Applied Psychology,* 73(4)**,** 703.

IUSSC (2009a). Departmental Report, Cmnd 7596, July. London:

IUSSC (2009b). *Students and Universities: Eleventh Report of Session 2008-09*. HC 170-1. London: HMSO.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning*, Springer.

Janis, I. L. (1982). *Groupthink: Psychological Studies of Policy Decisions and Fiascoes,* Boston, Houghton Mifflin.

Jenkins, S. (1995). *The Lady Who Turned to Nationalisation* [Online]. Times Higher Education Supplement. Available: http://www.timeshighereducation.co.uk/95716.article.

Jones, L., Allen, L., Abel, M., Abbott, S., Achinger, C., Addis, M., Alamino, R., Albiez, S., Anderson, G., Ansell, N., Barlow, T., Barnett, R., Bates, S., Bellamy, R., Bhambra, G. K., Bieler, A., Bishop, D., Bousquet, A., Brecher, B., Brett, P., Bulmer, S., Burkitt, I., Canaan, J., Cannell, F., Campbell, M., Castiglione, D., Clarke, D. I., Colquhoun, D., Cohen, R., Conway, D., Cowley, R., Cooke, B., Cross, H., Cruickshank, J., Curtis, D., Dendrinos, C., Dexter, H., Diaz, L., Dickey, E., Dirks, W., Docherty, T., Downer, J., Duxbury, C., Eichner, B., Edmond, N., Epstein, D., Evans, K., Faramelli, A., Arrigoitia, M. F., Fenwick, A., Filling, J., Frankham, J., Gilmore, J., G, M., Goetze, C., Goffey, A., Goodenough, J., Greenhalgh, T., Grivell, P.,

Harman, K., Hall, R., Hammersley, M., Hampshire, J., Haran, J., Hatcher, R., Hayes, D., Henderson, J., Hernández, Y., Hirst, A., Hoover, J., Holmwood, J., Holtschneider, H., Hopgood, S., Hotson, H., Houston, M., Humphreys, A., Iannou, I., Ingold, J., Inglis, F., Jarvis, S., John, E., Jones, C., Jones, S., Jesson, S., Khalili, L., Kirby, P., Kitchen, N., Krishnan, S., Kröger, S., Ladyman, J., Lawrence, E., Leitmeir, C., Levidow, L., Loughlin, M., Lyon, D., Maclean, M., Madden, M., Maisuria, A., Maunder, C., May, T., et al. 2015. *RE: The Use of 'Student Outcomes' to Measure Teaching Quality Is Completely Inappropriate.* Madeleine Atkins & Goodfellow, J.

JPG (1996a). Final Report. CVCP.

JPG (1996b). First Report. CVCP.

Kahneman, D. (2011). *Thinking, Fast and Slow*, Macmillan.

Kahneman, D. & Klein, G. (2009). Conditions for Intuitive Expertise: A Failure to Disagree. *American Psychologist,* 64(6)**,** 515.

Kahneman, D., Slovic, P. & Tversky, A. (eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases,* United States of America: Cambridge University Press.

Kavlock, R., Chandler, K., Houck, K., Hunter, S., Judson, R., Kleinstreuer, N., Knudsen, T., Martin, M., Padilla, S. & Reif, D. (2012). Update on Epa's Toxcast Program: Providing High Throughput Decision Support Tools for Chemical Risk Management. *Chemical research in toxicology,* 25(7)**,** 1287-1302.

Kewin, J. & Janowski, L. (2014). *Summary: SFCA Funding Impact Survey 2014* [Online]. Sixth Form Colleges Association. Available: http://www.sixthformcolleges.org/sites/default/files/160614%20SFCA%20Funding%20Impact%20Survey%20FINAL_0.pdf.

Kimber, I. (2015). *Metrics and Quality: Do the Numbers Add Up?* [Online]. Available: http://wonkhe.com/blogs/metrics-and-quality-do-the-numbers-add-up/.

King, R. (2011a). *The Risks of Risk-Based Regulation: The Regulatory Challenges of the Higher Education White Paper for England*, HEPI.

King, R. (2011b). *Talking About Quality: Comments on Colin Raban's Paper, Risk and Regulation* [Online]. Available: http://www.qaa.ac.uk/Publications/InformationAndGuidance/Documents/Roger-King-comment.pdf.

King, R. (2014a). *Regulating Uncertainty: Pluralism and Centralism in the Regulatory Regime for Higher Education in England* [Online]. Gloucester: Quality Assurance Agency. Available: http://www.qaa.ac.uk/Publications/InformationAndGuidance/Documents/Regulating-uncertainty.pdf.

King, R. (2014b). Risky Business: Academic Capitalism, Globalization, and the Risk University.

Kuhn, M. (2008). Building Predictive Models in R Using the Caret Package. *Journal of Statistical Software,* 28(5)**,** 1-26.

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*, Springer.

Lancaster, T. & Fanshawe, T. R. (2015). *Assessing the Quality of UK Medical Schools: What Is the Validity of Student Satisfaction Ratings as an Outcome Measure?* [Online]. Available: https://weblearn.ox.ac.uk/access/content/group/895746d8-7afb-4cb3-a461-7e7fca72e67c/Validity%20of%20student%20satisfaction%20ratings%20as%20a%20quality%20metric%20in%20UK%20medical%20education.pdf.

Lander, J. P. (2014). *R for Everyone: Advanced Analytics and Graphics*, Pearson Education.

Langlands, A. 6 Nov 2012. *RE: Invitation to the Quality Assurance Agency to Implement a More Risk-Based Approach to the Quality Assurance of Higher Education in England.* Letter to Mclaren, A.

Laplace, P-S. (1814). *A Philosophical Essay on Probabilities*.

Leach, M. (2016). *Universities Move to the Department for Education* [Online]. WonkHE. Available: http://wonkhe.com/blogs/higher-education-in-the-new-government/.

Lee, K. L., Pryor, D. B., Harrell, F. E., Califf, R. M., Behar, V. S., Floyd, W. L., Morris, J. J., Waugh, R. A., Whalen, R. E. & Rosati, R. A. (1986). Predicting Outcome in Coronary Disease Statistical Models Versus Expert Clinicians. *The American journal of medicine,* 80(4)**,** 553-560.

Leveson, B. (2012). Report of an Inquiry into the Culture, Practice and Ethics of the Press (the Stationery Office, 2012).

Liao, T. F. (1994). *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*, Sage.

Libby, R. (1976). Man Versus Model of Man: Some Conflicting Evidence. *Organizational Behavior and Human Performance,* 16(1)**,** 1-12.

Lind, S. (2014). *Gps in Shock over CQC 'Risk' Ratings* [Online]. Pulse. Available: http://www.pulsetoday.co.uk/your-practice/regulation/cqc/gps-in-shock-over-cqc-risk-ratings/20008577.fullarticle.

Lloyd's. (2015). *Lloyd's Minimum Standards* [Online]. Available: http://www.lloyds.com/~/media/files/the%20market/operating%20at%20lloyds/minimum%20standards/minimum%20standards%202/lloyds%20minimum%20standards.pdf.

Lloyd-Bostock, S. M. & Hutter, B. M. (2008). Reforming Regulation of the Medical Profession: The Risks of Risk-Based Approaches. *Health, Risk & Society,* 10(1)**,** 69-83.

Locke, K., Golden-Biddle, K. & Feldman, M. S. (2008). Perspective-Making Doubt Generative: Rethinking the Role of Doubt in the Research Process. *Organization Science,* 19(6)**,** 907-918.

Lodge, M. & Wegrich, K. (2012). *Managing Regulation: Regulatory Analysis, Politics and Policy*, Palgrave Macmillan.

Löfstedt, R. (2008). *Risk Management in Post-Trust Societies,* London, Earthscan.

Loughlin, M. & Scott, C. (1997). The Regulatory State. *In:* P. Dunleavy, A. Gamble, R. Heffernan & Peele, G. (eds.) *Developments in British Politics 5.* New York: St Martin's Press.

Mackinnon, I. & Norfolk, A. (2004). British University with Branch in Israeli Petrol Station 'Issued 5,500 Bogus Degrees. *The Times*, 22 Jan.

Majone, G. (1990). *Deregulation or Re-Regulation?: Regulatory Reform in Europe and the United States*, Pinter London.

Majone, G. (1994). The Rise of the Regulatory State in Europe. *West European Politics,* 17(3)**,** 77-101.

Majone, G. (1997). From the Positive to the Regulatory State: Causes and Consequences of Changes in the Mode of Governance. *Journal of public policy,* 17(02)**,** 139-167.

Martin, J. & Stephenson, R. (2005). Risk-Based Collection Model Development and Testing. *Internal Revenue Service Research Conference 2005.*

Martin, J. K. & Hirschberg, D. S. (1996). Small Sample Statistics for Classification Error Rates Ii: Confidence Intervals and Significance Tests. Information and Computer Science, University of California, Irvine

Mcclaren, A. & Brown, R. (2013). *New Arrangements for Quality Assurance in Higher Education* [Online]. HEPI. Available: http://www.hepi.ac.uk/wp-content/uploads/2014/02/HEPI-Occasional-Report-6-Quality-Assurance-in-Higher-Education-full-report.pdf.

Mcgettigan, A. (2013). *The Great University Gamble: Money, Markets and the Future of Higher Education*, PlutoPress.

Mcgettigan, A. (2014). Uncontrolled Expansion: How Private Colleges Grew. *Times Higher Edication*, 30th October, p.2170.

Mcknight, P. E., Mcknight, K. M., Sidani, S. & Figueredo, A. J. (2007). *Missing Data: A Gentle Introduction*, Guilford Press.

Meehl, P. E. (1954). Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.

Meehl, P. E. (1986). Causes and Effects of My Disturbing Little Book. *Journal of personality assessment,* 50(3)**,** 370-375.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L. & Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science,* 10(3)**,** 267-281.

Meyer, J. W. & Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony. *American journal of sociology,* 340-363.

Mickey, R. M. & Greenland, S. (1989). The Impact of Confounder Selection Criteria on Effect Estimation. *American journal of epidemiology,* 129(1)**,** 125-137.

Miller, G. A. (1962). *Psychology: The Science of Mental Life,* London, Penguin.

Miller, M. & Morris, N. (1988). Predictions of Dangerousness: An Argument for Limited Use. *Violence and victims,* 3(4)**,** 263-283.

Millett, D. & Bostock, N. (2014). *Exclusive: Gps Demand CQC Compensation* [Online]. GP Online. Available: http://www.gponline.com/exclusive-gps-demand-cqc-compensation/article/1325269.

Million+. (2014). *Million+ Comment on HEFCE Review of Higher Education Quality Assurance Arrangements* [Online]. Available: http://www.millionplus.ac.uk/press-releases/latest-press-releases/million-comment-on-hefce-review-of-higher-education-quality-assurance-arrangements.

Molinaro, A. M., Simon, R. & Pfeiffer, R. M. (2005). Prediction Error Estimation: A Comparison of Resampling Methods. *Bioinformatics,* 21(15)**,** 3301-3307.

Montier, J. (2009). *Behavioural Investing: A Practitioners Guide to Applying Behavioural Finance*, John Wiley & Sons.

MRUK Research. (2015). *The Future of Quality Assessment in Higher Education: Analysis of Responses to Phase 1 of the Quality Assessment Review* [Online]. HEFCE. Available: http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/Independentresearch/2015/The,future,of,QA,in,HE/2015_futureqainhe.pdf.

NAO (2000). *The Gaming Board: Better Regulation*. HC 537. London: The Stationer's Office.

NAO (2001). *Maritime and Coastal Agency: Ship Surveys and Inspections*. HC 338. London: NAO.

NAO (2002). *Tackling the Risks to Pension Scheme Members*. HC 1262. London: The Stationary Office.

NAO (2003). *Making a Difference: Performance of Maintained Secondary Schools in England*. HC 1332. London: The Stationary Office.

NAO (2009). *The Maritime and Coastguard Agency's Response to Growth in the UK Merchant Fleet*. HC 537. London: The Stationer's Office.

NAO (2014). *Investigation into Financial Support for Students at Alternative Higher Education Providers (Hc 861 Session 2014-15)* [Online]. Available: http://www.nao.org.uk/wp-content/uploads/2014/12/Investigation-into-financial-support-for-students-at-alternative-higher-education-providers.pdf.

Ncihe. (1997). *Higher Education in the Learning Society* [Online]. Available: http://www.leeds.ac.uk/educol/ncihe/.

NHS Improvement. (2015). *Board Meeting – 19 November 2015: Approach to Operational Planning for 2016/17* [Online]. Available: http://www.ntda.nhs.uk/wp-content/uploads/2015/09/Paper-J-Approach-to-operational-planning.pdf.

OECD (2010). *Risk and Regulatory Policy: Improving the Governance of Risk,* Paris, OECD.

OFFA (2015). *Facts and Figures on Tuition Fees and Student Finance* [Online]. Available: https://www.offa.org.uk/press/quick-facts/#key-facts.

Ofsted (2015a). *The Framework for School Inspection* [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/389974/The_framework_for_school_inspection.pdf.

Ofsted (2015b). *Methodology Note: The Risk Assessment of Good and Outstanding Providers* [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/448679/Methodology_note_the_risk_assessment_of_good_and_outstanding_providers.pdf.

ORR (2006). Train Derailment at Hatfield: A Final Report by the Independent Investigation Board. London: ORR

PA Consulting Group. (2000). *Better Accountability for Higher Education: Summary of a Review for the HEFCE (Report August 00/36)* [Online]. PA Consulting Group. Available: http://webarchive.nationalarchives.gov.uk/20100202100434/http://www.hefce.ac.uk/pubs/hefce/2000/00_36.pdf.

PAC (2014). *Monitor: Regulating NHS Foundation Trusts, Fourth Report of Session 2014–15*. HC 407, Incorporating HC 1119, Session 2013–14. London: The Stationary Office Limited.

PAC (2015). *Financial Support for Students at Alternative Higher Education Providers, Forty-First Report of Session 2014–15*. London: The Stationery Office Limited.

Park, M. Y. & Hastie, T. (2008). Penalized Logistic Regression for Detecting Gene Interactions. *Biostatistics,* 9(1)**,** 30-50.

Parker, C. (2002). *The Open Corporation: Effective Self-Regulation and Democracy*, Cambridge University Press.

Parry, G., Callender, C., Temple, P. & Scott, P. (2012). Understanding Higher Education in Further Education Colleges.

Parry, G. & Thompson, A. (2002). *Closer by Degrees: The Past, Present and Future of Higher Education in Further Education Colleges*, ERIC.

Patterson, L. & Lilburne, C. (2003). What Is the Commission for Health Improvement? *Postgraduate medical journal,* 79(932)**,** 303-305.

Peirce, C. S. (1958). *The Collect Papers of Charles Sanders Peirce,* Cambridge, MA, Harvard University Press.

Perrow, C. (1999). Normal Accidents, Updated Edition. Princeton University Press

Peters, T. J. & Waterman, R. H. (1982). *In Search of Excellence: Lessons from America's Best-Run Companies,* New York, Harper and Row.

Pfeffer, N. & Coote, A. (1991). *Is Quality Good for You?: A Critical Review of Quality Assurance in Welfare Services*, Institute for Public Policy Research.

Pirsig, R. M. (1974). *Zen and the Art of Motorcycle Maintenance,* New York, William Morrow and Company.

Pollard, A. (2011). *Witness Statement Provided for the Mid Staffordshire NHS Foundation Trust Public Inquiry, Monday 28 November* [Online]. Available: http://www.midstaffspublicinquiry.com/sites/default/files/evidence/Amanda_Pollard_-_witness_statement_and_exhibits.pdf.

Pollitt, C. (1993). Audit and Accountability: The Missing Dimension? *Journal of the Royal Society of Medicine,* 86(4)**,** 209.

Pollitt, C. (1995). Justification by Works or by Faith? Evaluating the New Public Management. *Evaluation,* 1(2)**,** 133-154.

Pontell, H. N. (1978). Deterrence Theory Versus Practice. *Criminology,* 16(1)**,** 3-22.

Power, M. (1997). The Audit Society: Rituals of Verification. *OUP Catalogue*.

Propper, C. & Wilson, D. (2003). The Use and Usefulness of Performance Measures in the Public Sector. *Oxford review of economic policy,* 19(2)**,** 250-267.

Provost, F. J., Fawcett, T. & Kohavi, R. (1998). Published. The Case against Accuracy Estimation for Comparing Induction Algorithms. ICML, 1998. 445-453.

QAA (1997). Work of the Agency. *Higher Quality,* 1.

QAA (1998a). An Agenda for Quality. *Higher Quality*, 3 Mar.

QAA (1998b). The Way Ahead. *Higher Quality*, 4 Oct.

QAA (2000). *Handbook for Academic Review,* Gloucester, QAA.

QAA (2002). *Handbook for Institutional Audit,* Gloucester, QAA.

QAA (2003). *Handbook for Enhancement Led Institutional Review,* Glasgow, QAA.

QAA (2005). *Learning from Developmental Engagements* [Online]. Quality Assurance Agency for Higher Education. Available: http://dera.ioe.ac.uk/8512/1/Learning%20from%20Development.pdf.

QAA (2010). *Evaluation of the Academic Infrastructure: Final Report* [Online]. Gloucester: QAA. Available: http://dera.ioe.ac.uk/1179/1/FinalReport.pdf.

QAA (2011a). *Institutional Review of Higher Education Institutions in England and Northern Ireland: A Handbook for Higher Education Providers,* Gloucester, QAA.

QAA (2011b). *The UK Quality Code for Higher Education* [Online]. Available: http://www.qaa.ac.uk/AssuringStandardsAndQuality/quality-code/Pages/default.aspx.

QAA (2011c). *The UK Quality Code for Higher Education - Part a: setting and maintaining academic Standards* [Online]. Available: http://www.qaa.ac.uk/assuring-standards-and-quality/the-quality-code/quality-code-part-a.

QAA (2012a). How Is Quality Assured?14/12/2015

QAA (2012b). *Protocol for Revisions to the UK Quality Code for Higher Education* [Online]. Available: http://www.qaa.ac.uk/en/Publications/Documents/Protocol-for-revisions-to-the-Quality-Code.pdf.

QAA (2012c). *Recognition Scheme for Educational Oversight: Handbook* [Online]. Available: http://www.qaa.ac.uk/en/Publications/Documents/RSEO-Handbook-2013.pdf.

QAA (2012d). *UK Quality Code for Higher Education - Chapter B1: Programme design, development And approval* [Online]. Gloucester: QAA. Available:

http://www.qaa.ac.uk/assuring-standards-and-quality/the-quality-code/quality-code-part-b.

QAA (2012e). *UK Quality Code for Higher Education - Chapter B2: recruitment, selection and Admission to higher education* [Online]. Gloucester: QAA. Available: http://www.qaa.ac.uk/assuring-standards-and-quality/the-quality-code/quality-code-part-b.

QAA (2012f). *UK Quality Code for Higher Education - Chapter B5: student Engagement* [Online]. Gloucester: QAA. Available: http://www.qaa.ac.uk/assuring-standards-and-quality/the-quality-code/quality-code-part-b.

QAA (2013a). *Higher Education Review: A Handbook for Providers* [Online]. Gloucester. Available: http://www.qaa.ac.uk/Publications/InformationAndGuidance/Documents/HER-handbook-13.pdf.

QAA (2013b). Outcomes of QAA Consultation on Higher Education Review

QAA (2013c). *UK Quality Code for Higher Education: General Introduction* [Online]. QAA. Available: http://www.qaa.ac.uk/en/Publications/Documents/Quality-Code-introduction.pdf.

QAA (2014a). *Audit of Tertiary Education Institutions: Mauritius* [Online]. Available: http://www.qaa.ac.uk/reviews-and-reports/audit-of-tertiary-education-institutions-mauritius.

QAA (2014b). *Higher Education Review (Plus): A Handbook for Alternative Providers Undergoing Review in 2014-15* [Online]. Gloucester. Available: http://www.qaa.ac.uk/en/Publications/Documents/HER-Plus-handbook-14.pdf.

QAA (2014c). *Higher Education Review: First Year Findings 2013-14* [Online]. Available: http://www.qaa.ac.uk/en/Publications/Documents/Findings-From-HER-2013-2014.pdf.

QAA (2014d). *Higher Education Review: What Happens after an Unsatisfactory Judgement?* [Online]. QAA. Available: http://www.qaa.ac.uk/en/Publications/Documents/HER-unsatisfactory-judgement.pdf.

QAA (2014e). *Review for Educational Oversight: Report of the Monitoring Visit of Bedfordian Business School, October 2014* [Online]. Available: http://www.qaa.ac.uk/en/ReviewsAndReports/Documents/Bedfordian%20Business%20School/Bedfordian-Business-School-REO-AM-14.pdf.

QAA (2014f). *UK Quality Code for Higher Education Part A: Setting and Maintaining Academic Standards. The Frameworks for Higher Education Qualifications of UK Degree-Awarding Bodies* [Online]. Available: http://www.qaa.ac.uk/en/Publications/Documents/qualifications-frameworks.pdf.

QAA (2015a). *About Us* [Online]. Available: http://www.qaa.ac.uk/about-us.

QAA (2015b). *QAA Board of Directors Meeting 10 June 2015: Annual Plan and Budget 2015-16* [Online]. Gloucester: QAA. Available: http://www.qaa.ac.uk/en/AboutUs/Documents/QAA%20Board%20Meeting%20-%20June%202015/QAA-Board-Draft-Annual-Plan-and-Budget-2015-16-June-2015.pdf.

QAA (2016a). *QAA Reviews of Alternative Providers: Key Findings 2013-15,* Gloucester, QAA.

QAA (2016b). *QAA Selected as Preferred Bidder for Funding Bodies' Contracts* [Online]. QAA. Available: http://www.qaa.ac.uk/newsroom/qaa-selected-as-preferred-bidder-for-hefce-contracts#.V8XoxpgrKUk.

QAA & GOsC. (2011). *General Osteopathic Council Review of Osteopathic Courses and Course Providers: Handbook for Course Providers* [Online]. Available: http://www.qaa.ac.uk/en/Publications/Documents/GOsC-handbook-providers.pdf.

QARSG (2015a). *Future Approaches to Quality Assessment in England, Wales and Northern Ireland: Analysis of Responses to Consultation* [Online]. Available: http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/2015/201530/2015_30.pdf.

QARSG (2015b). *Future Approaches to Quality Assessment in England, Wales, and Northern Ireland - Consultation* [Online]. Available: http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/2015/201511/2015_11_.pdf.

QARSG (2015c). *The Future of Quality Assessment in Higher Education* [Online]. Available: http://www.hefce.ac.uk/media/hefce/content/whatwedo/learningandteaching/assuringquality/qareview/discussion/QAR_Discussion.pdf.

Raban, C. (2011). Talking About Quality: Risk and Regulation. *The Quality Assurance Agency for Higher Education, Gloucester*.

Rail Safety and Standards Board (2005). Formal Inquiry: Derailment of Train 1t60, 1245 Hrs Kings Cross to Kings Lynn at Potters Bar on 10 May 2002 (Fi2013/F). London: Rail Safety and Standards Board.18 March 2005

Randall, J. (2000). A Profession for the New Millennium? *In:* Scott, P. (ed.) *Higher Education Reformed.* London: Falmer Press.

Raschka, S. (2015). *Python Machine Learning,* Birmingham, Packt Publishing Ltd.

Ratcliffe, R. & Adams, R. (2013). Derby University Accused of Falsifying Data on Graduate Employment Rate. *The Guardian*, 01/10/2013.

RCGP (2015). *CQC Chief Inspector Steve Field Has 'Lost Confidence of Gps' and Is 'No Longer Viewed as Fair and Impartial'* [Online]. RGCP. Available: http://www.rcgp.org.uk/news/2015/december/cqc-chief-inspector-steve-field-has-lost-confidence-of-gps.aspx.

Reason, J. (1990). *Human Error*, Cambridge university press.

Reynolds, P. A. (1986). *Academic Standards in Universities,* London, CVCP.

Rothstein, H. (2013). *Exploring National Cultures of Risk Governance* [Online]. Risk & Regulation Magazine: LSE. Available: http://www.lse.ac.uk/accounting/CARR/publications/CARRmagR&R25-Rothstein.pdf.

Rothstein, H., Borraz, O. & Huber, M. (2011). From the 'Neurotic' to the 'Rationalising' State: Risk and the Limits of Governance. *In:* Meyer, C. & Franco, C. D. (eds.) *Forecasting, Warning, and Transnational Risks: Is Prevention Possible?* Basingstoke: Palgrave Macmillan.

Rothstein, H., Huber, M. & Gaskell, G. (2006a). A Theory of Risk Colonization: The Spiralling Regulatory Logics of Societal and Institutional Risk. *Economy and society,* 35(1)**,** 91-112.

Rothstein, H., Irving, P., Walden, T. & Yearsley, R. (2006b). The Risks of Risk-Based Regulation: Insights from the Environmental Policy Domain. *Environment International,* 32(8)**,** 1056-1065.

RSS (2016). *Response to the Department for Business Innovation and Skills' Technical Consultation (Year 2) on the Teaching Excellence Framework* [Online]. RSS. Available: http://www.rss.org.uk/Images/PDF/influencing-change/2016/RSS-response-to-BIS-Technical-Consultation-on-Teaching-Excellence-Framework-year-2.pdf.

Schmitz, C. C. (1993). Assessing the Validity of Higher Education Indicators. *Journal of Higher Education,* 503-521.

Schutt, R. & O'neil, C. (2013). *Doing Data Science: Straight Talk from the Frontline*, O'Reilly Media, Inc.

Scott, J. (2015). Oral Evidence to the BIS Select Committee 'Assessing Quality in Higher Education' Inquiry. London: The Stationary Office Limited.Tuesday 1 December 2015

Shephard, G. (1995). Developing Quality Assurance in Partnership with the Institutions of Higher Education, Letter to Professor Gareth Roberts, Chair of Cvcp, from Rt Hon Gillian Shephard Mp, 21 Sep.

Shore, C. & Wright, S. (1999). Audit Culture and Anthropology: Neo-Liberalism in British Higher Education. *The Journal of the Royal Anthropological Institute,* 5(4)**,** 557-575.

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't,* London, Penguin.

SLC (2015). *Student Support for Higher Education in England* [Online]. Available: http://www.slc.co.uk/official-statistics/full-catalogue-of-official-statistics/student-support-for-higher-education-in-england.aspx.

Slovic, P. (1992). Perception of Risk: Reflections on the Psychometric Paradigm. *In:* Krimsky, S. & Golding, D. (eds.) *The Social Theories of Risk.* Westport, Connecticut: Praeger.

Smith, P. (1995). On the Unintended Consequences of Publishing Performance Data in the Public Sector. *International Journal of Public Administration,* 18(2-3)**,** 277-310.

Song, C., Boulier, B. L. & Stekler, H. O. (2007). The Comparative Accuracy of Judgmental and Model Forecasts of American Football Games. *International Journal of Forecasting,* 23(3)**,** 405-413.

Spiegelhalter, D. J. (2005). Funnel Plots for Comparing Institutional Performance. *Statistics in medicine,* 24(8)**,** 1185-1202.

SRA (2014). *SRA Regulatory Risk Framework* [Online]. Available: https://www.sra.org.uk/documents/solicitors/freedom-in-practice/risk-framework.pdf.

SRHE (1983). *The Leverhulme Report. Excellence in Diversity. Towards a New Strategy for Higher Education,* Guildford, SRHE.

SRHE (2015). *SRHE Annual Research Conference: Programme & Book of Abstracts (9-11 Dec 2015, Newport)* [Online]. SRHE. Available: https://www.srhe.ac.uk/conference2015/downloads/SRHE_Conf_2015_paper_summaries.pdf.

Starr, C. (1969). Social Benefit Versus Technological Risk. *Readings in Risk,* 183-194.

Stevens, R. (2005). *University to Uni: The Politics of Higher Education in England since 1944,* London, Politico's Publishing.

Storan, J. & Hudson, T. (2015). *Understanding College Higher Education - Literature Review* [Online]. Available: http://www.uel.ac.uk/wwwmedia/microsites/continuum/che/Draft-literature-review-(website).pdf.

Sunstein, C. R. (2002). *Risk and Reason: Safety, Law, and the Environment*, Cambridge University Press.

Tait, P., Pye, M. & Legard, J. (2008). Further Education and the Delivery of Higher-Level Qualifications: Understanding the Contribution of Further Education to the Delivery of Level 4 (Higher) and Professional Qualifications: Final Report.

Taleb, N. N. (2010). *The Black Swan: The Impact of the Highly Improbable Fragility*, Random House LLC.

TEQSA (2012). *Regulatory Risk Framework* [Online]. Available: http://teqsa.gov.au/sites/default/files/TEQSARegulatoryRiskFramework_0.pdf.

TEQSA (2013). *Media Release: Teqsa Welcomes Red Tape Review* [Online]. Melbourne: TEQSA. Available: http://www.teqsa.gov.au/sites/default/files/TEQSAWelcomesReview_MR29052013.pdf.

TEQSA (2014a). *Regulatory Approach* [Online]. Available: http://www.teqsa.gov.au/regulatory-approach.

TEQSA (2014b). *Risk Assessment Framework: Version 2.0* [Online]. Available: http://www.teqsa.gov.au/sites/default/files/publication-documents/TEQSARiskAssessFramework2014_0.pdf.

TEQSA (2015). *A Risk and Standards Based Approach to Quality Assurance in Australia's Diverse Higher Education Sector* [Online]. Available: http://www.teqsa.gov.au/sites/default/files/publication-documents/RiskStandardsSectorPaperFeb2015.pdf.

Tetlock, P. & Gardner, D. (2016). *Superforecasting: The Art and Science of Prediction*, Random House.

Thatcher, M. (2002). Delegation to Independent Regulatory Agencies: Pressures, Functions and Contextual Mediation. *West European Politics,* 25(1)**,** 125-147.

THE (1992). Quality Assurance Arrangements Are Going Wrong. *THE*.

THE (1993a). Student Hardship Is Hitting Standards. *THE*.

THE (1993b). V-Cs Reject Quality Red Tape. *THE*.

THE (1993c). V-Cs Slam Red Tape. *THE*.

THE (2001a). Lobbying Secures Rethink on University Ombudsman. *THE*, 19 Oct 2001.

THE (2001b). LSE Leads Revolt against QAA. *THE*, 23 Mar 2001.

THE (2002). *Analysis: Salvaged Ship Sets Sail out of the Storm, 8 Nov* [Online]. THE. Available: http://www.timeshighereducation.co.uk/news/analysis-salvaged-ship-sets-sail-out-of-the-storm/172819.article.

THE (2005a). *De Montfort in Exam Furore, 18 Mar* [Online]. THE. Available: http://www.timeshighereducation.co.uk/news/de-montfort-in-exam-furore/194764.article.

THE (2005b). *Luton Waived Entry Criteria, 14 Jan* [Online]. THE. Available: http://www.timeshighereducation.co.uk/news/luton-waived-entry-criteria/193410.article.

THE (2005c). *QAA Tells Leeds Met to Close Loophole, 6 May* [Online]. THE. Available: http://www.timeshighereducation.co.uk/news/qaa-tells-leeds-met-to-close-loophole/195831.article.

The Guardian. (2013). *University Guide 2014: University League Table* [Online]. Available: http://www.theguardian.com/education/table/2013/jun/03/university-league-table-2014.

Toft, B. (1996). Limits to the Mathematical Modelling of Disasters. *In:* Hood, C. & Jones, D. (eds.) *Accident and Design: Contemporary Debates in Risk Management.* London: UCL Press.

Turner, B. A. (1994). The Future for Risk Research. *Journal of Contingencies and Crisis Management,* 2(3)146-156.

Tynan, B. (2015). Oral Evidence to the BIS Select Committee 'Assessing Quality in Higher Education' Inquiry. London: The Stationary Office Limited.Tuesday 17 November 2015

UCAS (2015). *Foundation Degree Course Search* [Online]. Available: http://fd.ucas.com/FoundationDegree/About.aspx#q1.

UFC (1991). The 1992 Research Assessment Exercise, Circular Letter 22/91. 4 Oct 1991.

Universities UK. (2013). *Where Student Fees Go* [Online]. Available: http://www.universitiesuk.ac.uk/highereducation/Documents/2013/WhereStudentFeesGo.pdf.

Universities UK. (2015). *Quality, Equity, Sustainability: The Future of Higher Education Regulation* [Online]. Available: http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2015/quality-equity-sustainability.pdf.

Usher, R. (1997). Introduction. *In:* Mckenzie, G. W., Powell, J. & Usher, R. (eds.) *Understanding Social Research: Perspectives on Methodology and Practice.* London: Psychology Press.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*, CRC press.

Verostek, J. (2014). *Evaluating Model Performance* [Online]. Available: http://www.johnverostek.com/wp-content/uploads/2014/06/Chapter-10.pdf.

Vickers, J. (2011). Can Induction Be Justified? *In:* Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy.* Stanford: The Metaphysics Research Lab.

Wagner, L. (1993). The Teaching Quality Debate. *Higher Education Quarterly,* 47(3)274-85.

Walden, G. (1996). *We Should Know Better: Solving the Education Crisis*, Fourth Estate London.

Watson, D. (1995). Quality Assessment and 'Self-Regulation': The English Experience, 1992-94. *Higher Education Quarterly,* 49(4)326-40.

Watson, D. & Bowden, R. (1997). *Ends without Means: The Conservative Stewardship of UK Higher Education 1979-1997*, University of Brighton, Education Research Centre.

Wedding, D. (1983). Clinical and Statistical Prediction in Neuropsychology. *Clinical Neuropsychology*.

Wei, T. (2013). *Package 'Corrplot'* [Online]. The Comprehensive R Archive Network. Available: https://cran.r-project.org/web/packages/corrplot/corrplot.pdf.

Weick, K. E. (1976). Educational Organizations as Loosely Coupled Systems. *Administrative science quarterly,* 1-19.

Werner, P. D., Rose, T. L. & Yesavage, J. A. (1983). Reliability, Accuracy, and Decision-Making Strategy in Clinical Predictions of Imminent Dangerousness. *Journal of Consulting and Clinical Psychology,* 51(6)**,** 815.

Williams, P. (2009). Oral Evidence to the Innovation, Universities, Science and Skills Committee Students and Universities Inquiry. *Eleventh Report of Session 2008-09.* London: The Stationary Office Limited

Wilsdon, J. (2015a). *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management*, SAGE.

Wilsdon, J. (2015b). Oral Evidence to the BIS Select Committee 'Assessing Quality in Higher Education' Inquiry. London: The Stationary Office Limited.Tuesday 1 December 2015

Wolf, A. (1998). Two Sides of A4 Will Not Do the Trick. *THE*, 22 May.

Wolf, A. (2015). *Heading for the Precipice: Can Further and Higher Education Funding Policies Be Sustained?* [Online]. The Policy Institute at King's. Available: http://www.kcl.ac.uk/sspp/policy-institute/publications/Issuesandideas-alison-wolf-digital.pdf.

WonkHE (2015). *A Critical Moment for the Quality Debate* [Online]. Available: http://wonkhe.com/blogs/a-critical-moment-for-the-quality-debate/.

WonkHE (2016). *Peace in Our Time? The Other Future of Quality Assessment* [Online]. Available: http://wonkhe.com/blogs/quality-wars-peace/.

Yeung, K. (2016). Algorithmic Regulation and Intelligent Enforcement. *LSE CARR Workshop: 'Regulatory Scholarship in Crisis'.* London.20-21 June 2016

Zhao, Y. & Cen, Y. (2013). *Data Mining Applications with R*, Academic Press.